

PROJECTION PURSUIT THROUGH RELATIVE ENTROPY MINIMISATION

Jacques Touboul

Laboratoire de Statistique Théorique et Appliquée

Université Pierre et Marie Curie

jack_touboul@hotmail.com

Projection Pursuit methodology permits to solve the difficult problem of finding an estimate of a density defined on a set of very large dimension. In his seminal article, Huber (see "Projection Pursuit", Annals of Statistics, 1985) evidences the interest of the Projection Pursuit method thanks to the factorisation of a density into a Gaussian component and some residual density in a context of Kullback-Leibler divergence maximisation.

In the present article, we introduce a new algorithm, and in particular a test for the factorisation of a density estimated from an iid sample.

Keywords: Projection pursuit; Minimum Kullback-Leibler divergence maximisation; Robustness; Elliptical distribution

MSC(2000): 62H40 62G07 62G20 62H11.

1. Outline of the article

Projection Pursuit aims at creating one or several projections delivering a maximum of information on the structure of a data set irrespective of its size. Once a structure has been evidenced, the corresponding data are transformed through a Gaussianization. Recursively, this process is repeated in order to determine another structure in the remaining data until no further structure can be highlighted eventually. These kind of approaches for isolating structures were first studied by Friedman [Frie84] and Huber [HUB85]. Each of them details, through two different methodologies each, how to isolate such a structure and therefore how to estimate the density of the corresponding data.

However, since Mu Zhu [ZMU04] showed the two methodologies described by each of the above authors did not in fact turn out to be equivalent when the number of iterations in the algorithms exceeds the dimension of the space containing the data, we will consequently only concentrate on Huber's study while taking into account Mu Zhu's input.

After providing a brief overview of Huber's methodologies, we will then expose our approach and objective.

1.1. Huber's analytic approach

A density f on \mathbb{R}^d is considered. We then define an instrumental density g with the same mean and variance as f . According to Huber's approach, we first carry out the $K(f, g) = 0$ test - with K being the relative entropy (also called the Kullback-Leibler divergence). If the test is passed, then $f = g$ and the algorithm stops. If the test were not to be verified, based on the maximisation of $a \mapsto K(f_a, g_a)$ since $K(f, g) = K(f_a, g_a) + K(f \frac{g_a}{f_a}, g)$ and assuming that $K(f, g)$ is finite, Huber's methodology requires as a first step to define a vector a_1 and a density $f^{(1)}$ with

$$a_1 = \arg \inf_{a \in \mathbb{R}_*^d} K(f \frac{g_a}{f_a}, g) \text{ and } f^{(1)} = f \frac{g_{a_1}}{f_{a_1}}, \quad (1)$$

where \mathbb{R}_*^d is the set of non null vectors of \mathbb{R}^d and f_a (resp. g_a) represents the density of $a^\top X$ (resp. $a^\top Y$) when f (resp. g) is the density of X (resp. Y).

As a second step, Huber's algorithm replaces f with $f^{(1)}$ and repeats the first step.

Finally, a sequence (a_1, a_2, \dots) of vectors of \mathbb{R}_*^d and a sequence of densities $f^{(i)}$ are derived from the iterations of this process.

Remark 1

The algorithm enables us to generate a product approximation and even a product representation of f . Indeed, two rules can trigger the end of the process. The first one is the nullity of the relative entropy and the second one is the process reaching the d^{th} iteration. When these two rules are satisfied, the algorithm produces a product approximation of f . When only the first rule is satisfied, the algorithm generates a product representation of f .

Mathematically, for any integer j , such that $K(f^{(j)}, g) = 0$ with $j \leq d$, the process infers $f^{(j)} = g$, i.e. $f = g \prod_{i=1}^j \frac{f_{a_i}^{(i-1)}}{g_{a_i}}$ since by induction $f^{(j)} = f \prod_{i=1}^j \frac{g_{a_i}}{f_{a_i}^{(i-1)}}$. Likewise, when, for all j , it gets $K(f^{(j)}, g) > 0$

with $j \leq d$, it is assumed $g = f^{(d)}$ in order to obtain $f = g \prod_{i=1}^d \frac{f_{a_i}^{(i-1)}}{g_{a_i}}$, i.e. we approximate f with the product $g \prod_{i=1}^d \frac{f_{a_i}^{(i-1)}}{g_{a_i}}$.

Even if the condition $j \leq d$ is not met, the algorithm can also stop if the Kullback-Leibler divergence equals zero. Therefore, since by induction we have $f^{(j)} = f \prod_{i=1}^j \frac{g_{a_i}}{f_{a_i}^{(i-1)}}$ with $f^{(0)} = f$, we infer $g =$

$f \prod_{i=1}^j \frac{g_{a_i}}{f_{a_i}^{(i-1)}}$. We can thus represent f as $f = g \prod_{i=1}^j \frac{f_{a_i}^{(i-1)}}{g_{a_i}}$.

Finally, we remark that the algorithm implies that the sequence $(K(f^{(j)}, g))_j$ is decreasing and non negative with $f^{(0)} = f$.

1.2. Huber's synthetic approach

Maintaining the notations of the above section, we begin with performing the $K(f, g) = 0$ test; If the test is passed, then $f = g$ and the algorithm stops, otherwise, based on the maximisation of $a \mapsto K(f_a, g_a)$ since $K(f, g) = K(f_a, g_a) + K(f, g \frac{f_a}{g_a})$ and assuming that $K(f, g)$ is finite, Huber's methodology requires as a first step to define a vector a_1 and a density $g^{(1)}$ with

$$a_1 = \arg \inf_{a \in \mathbb{R}_*^d} K(f, g \frac{f_a}{g_a}) \text{ and } g^{(1)} = g \frac{f_{a_1}}{g_{a_1}}. \quad (2)$$

As a second step, Huber's algorithm replaces g with $g^{(1)}$ and repeats the first step.

Finally, a sequence (a_1, a_2, \dots) of vectors of \mathbb{R}_*^d and a sequence of densities $g^{(i)}$ are derived from the iterations of this process.

Remark 2

Similarly as in the analytic approach, this methodology allows us to generate a product approximation and even a product representation of f from g . Moreover, it also offers the same end of process rules. In other words, if for any j , such that $j \leq d$, we have $K(f, g^{(j)}) > 0$, then f is approximated with $g^{(d)}$. And if there exists j , such that $K(f, g^{(j)}) = 0$, then it holds $g^{(j)} = f$, i.e. f is represented by $g^{(j)}$. In this case, the relationship $K(f, g^{(j)}) = 0$ implies that $g^{(j)} = f$, i.e. since by induction we have $g^{(j)} = g \prod_{i=1}^j \frac{f_{a_i}}{g_{a_i}^{(i-1)}}$ with $g^{(0)} = g$, it holds $f = g \prod_{i=1}^j \frac{f_{a_i}}{g_{a_i}^{(i-1)}}$.

Eventually, we note that the algorithm implies that the sequence $(K(f, g^{(j)}))_j$ is decreasing and non negative with $g^{(0)} = g$.

Finally, in [ZMU04], Mu Zhu shows that, beyond d iterations, the data processing of these methodologies evidences significant differences, i.e. that past d iterations, the two methodologies are no longer equivalent. We will therefore only consider Huber's synthetic approach since g is known and since we want to find a representation of f .

1.3. Proposal

We begin with performing the $K(f, g) = 0$ test; should this test be passed, then $f = g$ and the algorithm stops, otherwise, the first step of our algorithm consists in defining a vector a_1 and a density $g^{(1)}$ by

$$a_1 = \arg \inf_{a \in \mathbb{R}_*^d} K(g \frac{f_a}{g_a}, f) \text{ and } g^{(1)} = g \frac{f_{a_1}}{g_{a_1}}. \quad (3)$$

In the second step, we replace g with $g^{(1)}$ and we repeat the first step. We thus derive, from the iterations of this process, a sequence (a_1, a_2, \dots) of vectors in \mathbb{R}_*^d and a sequence of densities $g^{(i)}$. We will prove that a_1 simultaneously optimises (1), (2) and (3). We will also prove that the underlying structures of f evidenced through this method are identical to the ones obtained through the Huber's methods.

Remark 3

As in Huber's algorithms, we perform a product approximation and even a product representation of f .

In the case where, at each of the d^{th} first steps, the relative entropy is positive, we then approximate f with $g^{(d)}$.

In the case where there exists a step of the algorithm such that the Kullback-Leibler divergence equals zero, then, calling j this step, we represent f with $g^{(dj)}$. In other words, if there exists a positive integer j such that $K(g^{(j)}, f) = 0$, then, since by induction we have $g^{(j)} = g \prod_{i=1}^j \frac{f_{a_i}}{g_{a_i}}$ with $g^{(0)} = g$, we represent f with the product $g \prod_{i=1}^j \frac{f_{a_i}}{g_{a_i}}$.

We also remark that the algorithm implies that the sequence $(K(g^{(j)}, f))_j$ is decreasing and non negative with $g^{(0)} = g$.

Finally, the very form of the relationship (3) demonstrates that we deal with M-estimation. We can consequently state that our method is more robust than Huber's - see [YOHAI], [TOMA] as well as [HUBER].

Example 1 *Let f be a density defined on \mathbb{R}^{10} by $f(x_1, \dots, x_{10}) = \eta(x_2, \dots, x_{10})\zeta(x_1)$, with η being a multivariate Gaussian density on \mathbb{R}^9 , and ζ being a non Gaussian density.*

Let us also consider g , a multivariate Gaussian density with the same mean and variance as f .

Since $g(x_2, \dots, x_{10}/x_1) = \eta(x_2, \dots, x_{10})$, we have $K(g \frac{f_1}{g_1}, f) = K(\eta \cdot f_1, f) = K(f, f) = 0$ as $f_1 = \zeta$ - where f_1 and g_1 are the first marginal densities of f and g respectively. Hence, the non negative function $a \mapsto K(g \frac{f_a}{g_a}, f)$ reaches zero for $e_1 = (1, 0, \dots, 0)'$.

We therefore infer that $g(x_2, \dots, x_{10}/x_1) = f(x_2, \dots, x_{10}/x_1)$.

To recapitulate our method, if $K(g, f) = 0$, we derive f from the relationship $f = g$; should a sequence $(a_i)_{i=1, \dots, j}$, $j < d$, of vectors in \mathbb{R}_*^d defining $g^{(j)}$ and such that $K(g^{(j)}, f) = 0$ exist, then $f(\cdot/a_i^\top x, 1 \leq i \leq j) = g(\cdot/a_i^\top x, 1 \leq i \leq j)$, i.e. f coincides with g on the complement of the vector subspace generated by the family $\{a_i\}_{i=1, \dots, j}$ - see also section 2.1.2. for details.

In this paper, after having clarified the choice of g , we will consider the statistical solution to the representation problem, assuming that f is unknown and X_1, X_2, \dots, X_m are i.i.d. with density f . We will provide asymptotic results pertaining to the family of optimizing vectors $a_{k,m}$ - that we will define more precisely below - as m goes to infinity. Our results also prove that the empirical representation scheme converges towards the theoretical one. Finally, we will compare Huber's optimisation methods with ours and we will present simulations.

2. The algorithm

2.1. The model

As described by Friedman [Frie84] and Diaconis [DIAFREE84], the choice of g depends on the family of distribution one wants to find in f . Until now, the choice has only been to use the class of Gaussian distributions. This can also be extended to the class of elliptical distributions.

2.1.1. Elliptical distributions

The fact that conditional densities with elliptical distributions are also elliptical - see [CAMBANIS81], [LANDS03] - enables us to use this class in our algorithm - and in Huber's algorithms.

Definition 1 X is said to abide by a multivariate elliptical distribution, denoted $X \sim E_d(\mu, \Sigma, \xi_d)$, if X has the following density, for any x in \mathbb{R}^d : $f_X(x) = \frac{c_d}{|\Sigma|^{1/2}} \xi_d\left(\frac{1}{2}(x - \mu)' \Sigma^{-1}(x - \mu)\right)$, where Σ is a $d \times d$ positive-definite matrix and where μ is an d -column vector, where ξ_d is referred as the "density generator", where c_d is a normalisation constant, such that $c_d = \frac{\Gamma(d/2)}{(2\pi)^{d/2}} \left(\int_0^\infty x^{d/2-1} \xi_d(x) dx \right)^{-1}$, with $\int_0^\infty x^{d/2-1} \xi_d(x) dx < \infty$.

Property 1 1/ For any $X \sim E_d(\mu, \Sigma, \xi_d)$, for any $m \times d$ matrix with rank $m \leq d$, A , and for any m -dimensional vector, b , we have $AX + b \sim E_m(A\mu + b, A\Sigma A', \xi_m)$.

Any marginal density of multivariate elliptical distribution is consequently elliptical, i.e.

$X = (X_1, X_2, \dots, X_d) \sim E_d(\mu, \Sigma, \xi_d)$ implies that $X_i \sim E_1(\mu_i, \sigma_i^2, \xi_1)$ with $f_{X_i}(x) = \frac{c_1}{\sigma_i} \xi_1\left(\frac{1}{2}\left(\frac{x - \mu_i}{\sigma_i}\right)^2\right)$, $1 \leq i \leq d$.

2/ Corollary 5 of [CAMBANIS81] states that the conditional densities with elliptical distributions are also elliptical. Indeed, if $X = (X_1, X_2)' \sim E_d(\mu, \Sigma, \xi_d)$, with X_1 (resp. X_2) of size $d_1 < d$ (resp. $d_2 < d$), then $X_1/(X_2 = a) \sim E_{d_1}(\mu', \Sigma', \xi_{d_1})$ with $\mu' = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$ and $\Sigma' = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, with $\mu = (\mu_1, \mu_2)$ and $\Sigma = (\Sigma_{ij})_{1 \leq i, j \leq 2}$.

Remark 4 In [LANDS03], the authors show that the multivariate Gaussian distribution derives from $\xi_d(x) = e^{-x}$. They also show that if $X = (X_1, \dots, X_d)$ has an elliptical density such that its marginals meet $E(X_i) < \infty$ and $E(X_i^2) < \infty$ for $1 \leq i \leq d$, then μ is the mean of X and Σ is a multiple of the covariance matrix of X . From now on, we will therefore assume this is the case.

Definition 2 Let t be an elliptical density on \mathbb{R}^k and let q be an elliptical density on $\mathbb{R}^{k'}$. The elliptical densities t and q are said to be part of the same family of elliptical densities, if their generating densities are ξ_k and $\xi_{k'}$ respectively, which belong to a common given family of densities.

Example 2 Consider two Gaussian densities $\mathcal{N}(0, 1)$ and $\mathcal{N}((0, 0), Id_2)$. They are said to belong to the same elliptical family as they both present $x \mapsto e^{-x}$ as generating density.

2.1.2. Choice of g

Let f be a density on \mathbb{R}^d . We assume there exists d non null linearly independent vectors a_j , with $1 \leq j \leq d$, of \mathbb{R}^d , such that

$$f(x) = n(a_{j+1}^\top x, \dots, a_d^\top x) h(a_1^\top x, \dots, a_j^\top x), \quad (4)$$

with $j < d$, n being an elliptical density on \mathbb{R}^{d-j-1} and with h being a density on \mathbb{R}^j , which does not belong to the same family as n . Let $X = (X_1, \dots, X_d)$ be a vector with f as density.

We define g as an elliptical distribution with the same mean and variance as f .

For simplicity, let us assume that the family $\{a_j\}_{1 \leq j \leq d}$ is the canonical basis of \mathbb{R}^d :

The very definition of f implies that (X_{j+1}, \dots, X_d) is independent from (X_1, \dots, X_j) . Hence, the property 1 allows us to infer that the density of (X_{j+1}, \dots, X_d) given (X_1, \dots, X_j) is n .

Let us assume that $K(g^{(j)}, f) = 0$, for some $j \leq d$. We then get $\frac{f(x)}{f_{a_1} f_{a_2} \dots f_{a_j}} = \frac{g(x)}{g_{a_1}^{(1-1)} g_{a_2}^{(2-1)} \dots g_{a_j}^{(j-1)}}$, since,

by induction, we have $g^{(j)}(x) = g(x) \frac{f_{a_1}}{g_{a_1}^{(1-1)}} \frac{f_{a_2}}{g_{a_2}^{(2-1)}} \dots \frac{f_{a_j}}{g_{a_j}^{(j-1)}}$. Consequently, the fact that the conditional densities with elliptical distributions are also elliptical, as well as the above relationship enable us to state that $n(a_{j+1}^\top x, \dots, a_d^\top x) = f(\cdot / a_i^\top x, 1 \leq i \leq j) = g(\cdot / a_i^\top x, 1 \leq i \leq j)$. In other words, f coincides with g on the complement of the vector subspace generated by the family $\{a_i\}_{i=1, \dots, j}$.

At present, if the family $\{a_j\}_{1 \leq j \leq d}$ is no longer the canonical basis of \mathbb{R}^d , then this family is again a basis of \mathbb{R}^d . Hence, lemma 11 implies that

$$g(\cdot / a_1^\top x, \dots, a_j^\top x) = n(a_{j+1}^\top x, \dots, a_d^\top x) = f(\cdot / a_1^\top x, \dots, a_j^\top x), \quad (5)$$

which is equivalent to $K(g^{(j)}, f) = 0$, since by induction $g^{(j)} = g \frac{f_{a_1}}{g_{a_1}^{(1-1)}} \frac{f_{a_2}}{g_{a_2}^{(2-1)}} \dots \frac{f_{a_j}}{g_{a_j}^{(j-1)}}$.

The end of our algorithm implies that f coincides with g on the complement of the vector subspace generated by the family $\{a_i\}_{i=1, \dots, j}$. Therefore, the nullity of the Kullback-Leibler divergence provides us with information on the density structure. In summary, the following proposition clarifies the choice of g which depends on the family of distribution one wants to find in f :

Proposition 1 *With the above notations, $K(g^{(j)}, f) = 0$ is equivalent to*

$$g(\cdot / a_1^\top x, \dots, a_j^\top x) = f(\cdot / a_1^\top x, \dots, a_j^\top x).$$

More generally, the above proposition leads us to defining the co-support of f as the vector space generated by the vectors a_1, \dots, a_j .

Definition 3 *Let f be a density on \mathbb{R}^d . We define the co-vectors of f as the sequence of vectors a_1, \dots, a_j which solves the problem $K(g^{(j)}, f) = 0$ where g is an elliptical distribution with the same mean and variance as f . We define the co-support of f as the vector space generated by the vectors a_1, \dots, a_j .*

2.2. Stochastic outline of the algorithm

Let X_1, X_2, \dots, X_m (resp. Y_1, Y_2, \dots, Y_m) be a sequence of m independent random vectors with the same density f (resp. g). As customary in nonparametric Kullback-Leibler optimizations, all estimates of f and f_a , as well as all uses of Monte Carlo methods are being performed using subsamples X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_n , extracted respectively from X_1, X_2, \dots, X_m and Y_1, Y_2, \dots, Y_m , since the estimates are bounded below by some positive deterministic sequence θ_m (see Appendix B).

Let \mathbb{P}_n be the empirical measure based on the subsample X_1, X_2, \dots, X_n . Let f_n (resp. $f_{a,n}$ for any a in \mathbb{R}_*^d) be the kernel estimate of f (resp. f_a), which is built from X_1, X_2, \dots, X_n (resp. $a^\top X_1, a^\top X_2, \dots, a^\top X_n$).

As defined in section 1.3, we introduce the following sequences $(a_k)_{k \geq 1}$ and $(g^{(k)})_{k \geq 1}$:

- a_k is a non null vector of \mathbb{R}^d such that $a_k = \arg \min_{a \in \mathbb{R}_*^d} K(g^{(k-1)} \frac{f_a}{g_a^{(k-1)}}, f)$,
- $g^{(k)}$ is the density such that $g^{(k)} = g^{(k-1)} \frac{f_{a_k}}{g_{a_k}^{(k-1)}}$ with $g^{(0)} = g$.

The stochastic setting up of the algorithm uses f_n and $g_n^{(0)} = g$ instead of f and $g^{(0)} = g$, since g is known. Thus, at the first step, we build the vector \check{a}_1 which minimizes the Kullback-Leibler divergence between f_n and $g \frac{f_{a,n}}{g_a}$ and which estimates a_1 .

Proposition 10 and lemma 12 enable us to minimize the Kullback-Leibler divergence between f_n and $g \frac{f_{a,n}}{g_a}$. Defining \check{a}_1 as the argument of this minimization, proposition 4 shows us that this vector tends to a_1 .

Finally, we define the density $\check{g}_m^{(1)}$ as $\check{g}_m^{(1)} = g \frac{f_{\check{a}_1, m}}{g_{\check{a}_1}^{(1)}}$ which estimates $g^{(1)}$ through theorem 1.

Now, from the second step and as defined in section 1.3, the density $g^{(k-1)}$ is unknown. Once again, we therefore have to truncate the samples.

All estimates of f and f_a (resp. $g^{(1)}$ and $g_a^{(1)}$) are being performed using a subsample X_1, X_2, \dots, X_n (resp. $Y_1^{(1)}, Y_2^{(1)}, \dots, Y_n^{(1)}$) extracted from X_1, X_2, \dots, X_m (resp. $Y_1^{(1)}, Y_2^{(1)}, \dots, Y_m^{(1)}$) - which is a sequence of m independent random vectors with the same density $g^{(1)}$ such that the estimates are bounded below by some positive deterministic sequence θ_m (see Appendix B).

Let \mathbb{P}_n be the empirical measure based on the subsample X_1, X_2, \dots, X_n . Let f_n (resp. $g_n^{(1)}, f_{a,n}, g_{a,n}^{(1)}$ for any a in \mathbb{R}_*^d) be the kernel estimate of f (resp. $g^{(1)}, f_a, g_a^{(1)}$) which is built from X_1, X_2, \dots, X_n (resp. $Y_1^{(1)}, Y_2^{(1)}, \dots, Y_n^{(1)}$). The stochastic setting up of the algorithm uses f_n and $g_n^{(1)}$ instead of f and $g^{(1)}$. Thus, we build the vector \check{a}_2 which minimizes the Kullback-Leibler divergence between f_n and $g_n^{(1)} \frac{f_{a,n}}{g_{a,n}^{(1)}}$ - since $g^{(1)}$ and $g_a^{(1)}$ are unknown - and which estimates a_2 . Proposition 10 and lemma 12 enable us to minimize the Kullback-Leibler divergence between f_n and $g_n^{(1)} \frac{f_{a,n}}{g_{a,n}^{(1)}}$. Defining \check{a}_2 as the argument of this minimization, proposition 4 shows that this vector tends to a_2 in n . Finally, we define the density $\check{g}_n^{(2)}$ as $\check{g}_n^{(2)} = g_n^{(1)} \frac{f_{\check{a}_2, n}}{g_{\check{a}_2, n}^{(1)}}$ which estimates $g^{(2)}$ through theorem 1.

And so on, we will end up obtaining a sequence $(\check{a}_1, \check{a}_2, \dots)$ of vectors in \mathbb{R}_*^d estimating the co-vectors of f and a sequence of densities $(\check{g}_n^{(k)})_k$ such that $\check{g}_n^{(k)}$ estimates $g^{(k)}$ through theorem 1.

3. Results

3.1. Convergence results

3.1.1. Hypotheses on f

In this paragraph, we define the set of hypotheses on f which can possibly be used in our work. Discussion on several of these hypotheses can be found in Appendix D. In this section, to be more legible we replace g with $g^{(k-1)}$. Let $\Theta = \mathbb{R}_*^d$, $M(b, a, x) = \int \ln\left(\frac{g(x)}{f(x)} \frac{f_b(b^\top x)}{g_b(b^\top x)}\right) g(x) \frac{f_a(a^\top x)}{g_a(a^\top x)} dx - \left(\frac{g(x)}{f(x)} \frac{f_b(b^\top x)}{g_b(b^\top x)} - 1\right)$,

$$\mathbb{P}_n M(b, a) = \int M(b, a, x) d\mathbb{P}_n, \quad \mathbf{P} M(b, a) = \int M(b, a, x) f(x) dx,$$

\mathbf{P} being the probability measure of f . Similarly as in chapter V of [VDW], we define :

(H'1) : For all $\varepsilon > 0$, there is $\eta > 0$, such that for all $c \in \Theta$ verifying

$$\|c - a_k\| \geq \varepsilon, \text{ we have } \mathbf{P} M(c, a) < \mathbf{P} M(a_k, a) - \eta, \text{ with } a \in \Theta.$$

(H'2) : There exists a neighborhood of a_k , V , and a positive function H , such

$$\text{that, for all } c \in V \text{ we have } |M(c, a_k, x)| \leq H(x) \text{ (}\mathbf{P} - a.s.) \text{ with } \mathbf{P} H < \infty,$$

(H'3) : There exists a neighborhood of a_k , V , such that for all ε , there exists a η such that

for all $c \in V$ and $a \in \Theta$, verifying $\|a - a_k\| \geq \varepsilon$, we have $\mathbf{P}M(c, a_k) < \mathbf{P}M(c, a) - \eta$.

Putting $I_{a_k} = \frac{\partial^2}{\partial a^2} K(g \frac{f_{a_k}}{g_{a_k}}, f)$, and $x \rightarrow \rho(b, a, x) = \ln \left(\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)} \right) \frac{g(x)f_a(a^\top x)}{g_a(a^\top x)}$, we now consider :

(H'4) : There exists a neighborhood of (a_k, a_k) , V'_k , such that, for all (b, a) of V'_k , the gradient $\nabla \left(\frac{g(x)f_a(a^\top x)}{g_a(a^\top x)} \right)$ and the Hessian $\mathcal{H} \left(\frac{g(x)f_a(a^\top x)}{g_a(a^\top x)} \right)$ exist (λ -a.s.), and the first order partial derivative $\frac{g(x)f_a(a^\top x)}{g_a(a^\top x)}$ and the first and second order derivative of $(b, a) \mapsto \rho(b, a, x)$ are dominated (λ -a.s.) by integrable functions.

(H'5) : The function $(b, a) \mapsto M(b, a, x)$ is \mathcal{C}^3 in a neighborhood V'_k of (a_k, a_k) for all x and all the partial derivatives of order 3 of $(b, a) \mapsto M(b, a, x)$ are dominated in V'_k by a \mathbf{P} -integrable function $H(x)$.

(H'6) : $\mathbf{P} \left\| \frac{\partial}{\partial b} M(a_k, a_k) \right\|^2$ and $\mathbf{P} \left\| \frac{\partial}{\partial a} M(a_k, a_k) \right\|^2$ are finite and the expressions $\mathbf{P} \frac{\partial^2}{\partial b_i \partial b_j} M(a_k, a_k)$ and I_{a_k} exist and are invertible.

(H'7) : There exists k such that $\mathbf{P}M(a_k, a_k) = 0$.

(H'8) : $(\text{Var}_{\mathbf{P}}(M(a_k, a_k)))^{1/2}$ exists and is invertible.

(H'0): f and g are assumed to be positive and bounded and such that $K(g, f) \geq \int |f(x) - g(x)| dx$.

3.1.2. Estimation of the first co-vector of f

Let \mathcal{R} be the class of all positive functions r defined on \mathbb{R} and such that $g(x)r(a^\top x)$ is a density on \mathbb{R}^d for all a belonging to \mathbb{R}_*^d . The following proposition shows that there exists a vector a such that $\frac{f_a}{g_a}$ minimizes $K(gr, f)$ in r :

Proposition 2 *There exists a vector a belonging to \mathbb{R}_*^d such that*

$$\arg \min_{r \in \mathcal{R}} K(gr, f) = \frac{f_a}{g_a} \text{ and } r(a^\top x) = \frac{f_a(a^\top x)}{g_a(a^\top x)}.$$

Following [BROKEZ], let us introduce the estimate of $K(g \frac{f_{a,n}}{g_a}, f_n)$, through

$$\check{K}(g \frac{f_{a,n}}{g_a}, f_n) = \int M(a, a, x) d\mathbb{P}_n(x)$$

Proposition 3 *Let $\check{a} := \arg \inf_{a \in \mathbb{R}_*^d} \check{K}(g \frac{f_{a,n}}{g_a}, f_n)$.*

Then, \check{a} is a strongly convergent estimate of a , as defined in proposition 2.

Let us also introduce the following sequences $(\check{a}_k)_{k \geq 1}$ and $(\check{g}_n^{(k)})_{k \geq 1}$, for any given n - see section 2.2.:

- \check{a}_k is an estimate of a_k as defined in proposition 3 with $\check{g}_n^{(k-1)}$ instead of g ,
- $\check{g}_n^{(k)}$ is such that $\check{g}_n^{(0)} = g$, $\check{g}_n^{(k)}(x) = \check{g}_n^{(k-1)}(x) \frac{f_{\check{a}_k, n}(\check{a}_k^\top x)}{[\check{g}_n^{(k-1)}]_{\check{a}_k, n}(\check{a}_k^\top x)}$, i.e. $\check{g}_n^{(k)}(x) = g(x) \prod_{j=1}^k \frac{f_{\check{a}_j, n}(\check{a}_j^\top x)}{[\check{g}_n^{(j-1)}]_{\check{a}_j, n}(\check{a}_j^\top x)}$.

We also note that $\check{g}_n^{(k)}$ is a density.

3.1.3. Convergence study at the k^{th} step of the algorithm:

In this paragraph, we show that the sequence $(\check{a}_k)_n$ converges towards a_k and that the sequence $(\check{g}_n^{(k)})_n$ converges towards $g^{(k)}$.

Let $\check{c}_n(a) = \arg \sup_{c \in \Theta} \mathbb{P}_n M(c, a)$, with $a \in \Theta$, and $\check{\gamma}_n = \arg \inf_{a \in \Theta} \sup_{c \in \Theta} \mathbb{P}_n M(c, a)$. We state

Proposition 4 *Both $\sup_{a \in \Theta} \|\check{c}_n(a) - a_k\|$ and $\check{\gamma}_n$ converge toward a_k a.s.*

Finally, the following theorem shows that $\check{g}_n^{(k)}$ converges almost everywhere towards $g^{(k)}$:

Theorem 1 *It holds $\check{g}_n^{(k)} \rightarrow_n g^{(k)}$ a.s.*

3.2. Asymptotic inference at the k^{th} step of the algorithm

The following theorem shows that $\check{g}_n^{(k)}$ converges towards $g^{(k)}$ at the rate $O_{\mathbf{P}}(m^{-\frac{1}{4+d}})$ in three different cases, namely for any given x , with the L^1 distance and with the Kullback-Leibler divergence:

Theorem 2 *It holds $|\check{g}_n^{(k)}(x) - g^{(k)}(x)| = O_{\mathbf{P}}(m^{-\frac{1}{4+d}})$, $\int |\check{g}_n^{(k)}(x) - g^{(k)}(x)| dx = O_{\mathbf{P}}(m^{-\frac{1}{4+d}})$ and $|K(\check{g}_n^{(k)}, f) - K(g^{(k)}, f)| = O_{\mathbf{P}}(m^{-\frac{1}{4+d}})$.*

Then, the following theorem shows that the laws of our estimators of a_k , namely $\check{c}_n(a_k)$ and $\check{\gamma}_n$, converge towards a linear combination of Gaussian variables.

Theorem 3 *It holds $\sqrt{n}\mathcal{A}(\check{c}_n(a_k) - a_k) \xrightarrow{\mathcal{L}^{aw}} \mathcal{B}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}M(a_k, a_k)\|^2) + \mathcal{C}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}M(a_k, a_k)\|^2)$ and $\sqrt{n}\mathcal{A}(\check{\gamma}_n - a_k) \xrightarrow{\mathcal{L}^{aw}} \mathcal{C}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}M(a_k, a_k)\|^2) + \mathcal{C}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}M(a_k, a_k)\|^2)$ where $\mathcal{A} = \mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k, a_k)(\mathbf{P}\frac{\partial^2}{\partial a\partial a}M(a_k, a_k) + \mathbf{P}\frac{\partial^2}{\partial a\partial b}M(a_k, a_k))$, $\mathcal{C} = \mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k, a_k)$ and $\mathcal{B} = \mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k, a_k) + \mathbf{P}\frac{\partial^2}{\partial a\partial a}M(a_k, a_k) + \mathbf{P}\frac{\partial^2}{\partial a\partial b}M(a_k, a_k)$.*

3.3. A stopping rule for the procedure

In this paragraph, we show that $g_n^{(k)}$ converges towards f in k and n . Then, we provide a stopping rule for this identification procedure.

3.3.1. Estimation of f

Through remark 5 and as explained in section 14 of [HUB85], the following lemma shows that $K(g_n^{(k-1)} \frac{f_{a_k, n}}{g_{a_k, n}^{(k-1)}}, f_{a_k, n})$ converges almost everywhere towards zero as k goes to infinity and thereafter as n goes to infinity :

Lemma 1 *We have $\lim_n \lim_k K(\check{g}_n^{(k)} \frac{f_{a_k, n}}{[\check{g}^{(k)}]_{a_k, n}}, f_n) = 0$ a.s.*

Consequently, the following proposition provides us with an estimate of f :

Theorem 4 *We have $\lim_n \lim_k \check{g}_n^{(k)} = f$ a.s.*

3.3.2. Testing of the criteria

In this paragraph, through a test of the criteria, namely $a \mapsto K(\check{g}_n^{(k)} \frac{f_{a, n}}{[\check{g}^{(k)}]_{a, n}}, f_n)$, we build a stopping rule for this identification procedure. First, the next theorem enables us to derive the law of the criteria:

Theorem 5 *For a fixed k , we have*

$$\sqrt{n}(\text{Var}_{\mathbf{P}}(M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n)))^{-1/2}(\mathbb{P}_n M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n) - \mathbb{P}_n M(a_k, a_k)) \xrightarrow{\mathcal{L}^{aw}} \mathcal{N}(0, I),$$

as n goes to infinity, where k represents the k^{th} step of the algorithm and I is the identity matrix in \mathbb{R}^d .

Note that k is fixed in theorem 5 since $\check{\gamma}_n = \arg \inf_{a \in \Theta} \sup_{c \in \Theta} \mathbb{P}_n M(c, a)$ where M is a known function of k , see section 3.1.1. Thus, in the case where $K(g^{(k-1)} \frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f) = 0$, we obtain

Corollary 1 *We have $\sqrt{n}(\text{Var}_{\mathbf{P}}(M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n)))^{-1/2}(\mathbb{P}_n M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n)) \xrightarrow{\mathcal{L}^{aw}} \mathcal{N}(0, I)$.*

Hence, we propose the test of the null hypothesis

$$(H_0) : K(g^{(k-1)} \frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f) = 0 \text{ versus } (H_1) : K(g^{(k-1)} \frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f) \neq 0.$$

Based on this result, we stop the algorithm, then, defining a_k as the last vector generated, we derive from corollary 1 a α -level confidence ellipsoid around a_k , namely

$$\mathcal{E}_k = \{b \in \mathbb{R}^d; \sqrt{n}(\text{Var}_{\mathbf{P}}(M(b, b)))^{-1/2} \mathbb{P}_n M(b, b) \leq q_{\alpha}^{\mathcal{N}(0,1)}\}$$

where $q_{\alpha}^{\mathcal{N}(0,1)}$ is the quantile of a α -level reduced centered normal distribution and where \mathbb{P}_n is the empirical measure arising from a realization of the sequences (X_1, \dots, X_n) and (Y_1, \dots, Y_n) .

The following corollary thus provides us with a confidence region for the above test:

Corollary 2 \mathcal{E}_k is a confidence region for the test of the null hypothesis (H_0) versus (H_1) .

4. Comparison of all the optimisation methods

In this section, we study Huber's algorithm in a similar manner to sections 2 and 3. We will then be able to compare our methodologies.

Until now, the choice has only been to use the class of Gaussian distributions. Here and similarly to section 2.1, we extend this choice to the class of elliptical distributions. Moreover, using the subsample X_1, X_2, \dots, X_n , see Appendix B, and using the procedure of section 2.2. with $K(g_a, f_a)$, see section 4.2, instead of $K(g \frac{g_a}{f_a}, f)$, proposition 10, lemma 12 and remark 5 enable us to perform the Huber's algorithm :

- we define \hat{a}_1 and the density $\hat{g}_n^{(1)}$ such that $\hat{a}_1 = \arg \max_{a \in \mathbb{R}_*^d} K(g_a, f_{a,n})$ and $\hat{g}_n^{(1)} = g \frac{f_{\hat{a}_1, n}}{g_{\hat{a}_1}^{(1)}}$,
- we define \hat{a}_2 and the density $\hat{g}_n^{(2)}$ such that $\hat{a}_2 = \arg \max_{a \in \mathbb{R}_*^d} K(\hat{g}_{a,n}^{(1)}, f_{a,n})$ and $\hat{g}_n^{(2)} = \hat{g}_n^{(1)} \frac{f_{\hat{a}_2, n}}{\hat{g}_{\hat{a}_2, n}^{(1)}}$,

and so on, we obtain a sequence $(\hat{a}_1, \hat{a}_2, \dots)$ of vectors in \mathbb{R}_*^d and a sequence of densities $\hat{g}_n^{(k)}$.

4.1. Hypotheses on f

In this paragraph, we define the set of hypotheses on f which can be of use in our present work.

First, we denote g in lieu of $g^{(k-1)}$. Let $\Theta_a^1 = \{b \in \Theta \mid \int (\frac{g_b(b^\top x)}{f_b(b^\top x)} - 1) f_a(a^\top x) dx < \infty\}$,

$$m(b, a, x) = \int \ln(\frac{g_b(b^\top x)}{f_b(b^\top x)}) g_a(a^\top x) dx - (\frac{g_b(b^\top x)}{f_b(b^\top x)} - 1),$$

$$\mathbf{P}^a m(b, a) = \int m(b, a, x) f_a(a^\top x) dx \text{ and } \mathbb{P}_n m(b, a) = \int m(b, a, x) \frac{f_a(a^\top x)}{f(x)} d\mathbb{P}_n,$$

\mathbf{P}^a being the probability measure of f_a .

Similarly as in chapter V of [VDW], we define :

(H1) : For all $\varepsilon > 0$, there is $\eta > 0$ such that, for all $b \in \Theta_a^1$ verifying

$$\|b - a_k\| \geq \varepsilon \text{ for all } a \in \Theta, \text{ we have } \mathbf{P}^a m(b, a) < \mathbf{P}^a m(a_k, a) - \eta,$$

(H2) : There exists a neighborhood of a_k , V , and a positive function H , such

$$\text{that, for all } b \in V, \text{ we have } |m(b, a_k, x)| \leq H(x) \text{ } (\mathbf{P}^a - a.s.) \text{ with } \mathbf{P}^a H < \infty,$$

(H3) : There exists a neighborhood V of a_k , such that for all ε , there exists a η such

$$\text{that for all } b \in V \text{ and } a \in \Theta, \text{ verifying } \|a - a_k\| \geq \varepsilon, \text{ we have } \mathbf{P}^{a_k} m(b, a_k) - \eta > \mathbf{P}^a m(b, a).$$

Moreover, defining $x \rightarrow v(b, a, x) = \ln(\frac{g_a(a^\top x)}{f_a(a^\top x)}) g_a(a^\top x)$, putting:

(H4) : There exists a neighborhood of (a_k, a_k) , V_k , such that, for all (b, a) of V_k ,

the gradient $\nabla(\frac{g_a(a^\top x)}{f_a(a^\top x)})$ and the Hessian $\mathcal{H}(\frac{g_a(a^\top x)}{f_a(a^\top x)})$ exist ($\lambda - a.s.$) and the first order partial derivative $\frac{g_a(a^\top x)}{f_a(a^\top x)}$ and the first and second order derivative of order 3 of $(b, a) \mapsto v(b, a, x)$ are dominated ($\lambda - a.s.$) by integrable functions.

(H5) : The function $(b, a) \mapsto m(b, a)$ is \mathcal{C}^3 in a neighborhood V_k of (a_k, a_k) for all x and all the partial derivatives of $(b, a) \mapsto m(b, a)$ are dominated in V_k by a \mathbf{P} -integrable function $H(x)$.

(H6) : $\mathbf{P} \|\frac{\partial}{\partial b} m(a_k, a_k)\|^2$ and $\mathbf{P} \|\frac{\partial}{\partial a} m(a_k, a_k)\|^2$ are finite and the quantities

$\mathbf{P} \frac{\partial^2}{\partial b_i \partial b_j} m(a_k, a_k)$ and $\mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} m(a_k, a_k)$ are invertible.

(H7) : There exists k such that $\mathbf{P} m(a_k, a_k) = 0$.

(H8) : $(\text{Var}_{\mathbf{P}}(m(a_k, a_k)))^{1/2}$ exists and is invertible.

4.2. The first co-vector of f simultaneously optimizes four problems

We first study Huber's analytic approach. Let \mathcal{R}' be the class of all positive functions r defined on \mathbb{R} and such that $f(x)r^{-1}(a^\top x)$ is a density on \mathbb{R}^d for all a belonging to \mathbb{R}_*^d . The following proposition shows that there exists a vector a such that $\frac{f_a}{g_a}$ minimizes $K(fr^{-1}, g)$ in r :

Proposition 5 (Analytic Approach) *There exists a vector a belonging to \mathbb{R}_*^d such that $\arg \min_{r \in \mathcal{R}'} K(fr^{-1}, g) = \frac{f_a}{g_a}$, $r(a^\top x) = \frac{f_a(a^\top x)}{g_a(a^\top x)}$ as well as $K(f, g) = K(f_a, g_a) + K(f \frac{g_a}{f_a}, g)$.*

We also study Huber's synthetic approach. Let \mathcal{R} be the class of all positive functions r defined on \mathbb{R} and such that $g(x)r(a^\top x)$ is a density on \mathbb{R}^d for all a belonging to \mathbb{R}_*^d . The following proposition shows that there exists a vector a such that $\frac{f_a}{g_a}$ minimizes $K(gr, f)$ in r :

Proposition 6 (Synthetic Approach) *There exists a vector a belonging to \mathbb{R}_*^d such that $\arg \min_{r \in \mathcal{R}} K(f, gr) = \frac{f_a}{g_a}$, $r(a^\top x) = \frac{f_a(a^\top x)}{g_a(a^\top x)}$ as well as $K(f, g) = K(f_a, g_a) + K(f, g \frac{f_a}{g_a})$.*

In the meanwhile, the following proposition shows that there exists a vector a such that $\frac{f_a}{g_a}$ minimizes $K(g, fr^{-1})$ in r .

Proposition 7 *There exists a vector a belonging to \mathbb{R}_*^d such that $\arg \min_{r \in \mathcal{R}'} K(g, fr^{-1}) = \frac{f_a}{g_a}$, and $r(a^\top x) = \frac{f_a(a^\top x)}{g_a(a^\top x)}$. Moreover, we have $K(g, f) = K(g_a, f_a) + K(g, f \frac{g_a}{f_a})$.*

Remark 5 *First, through property 4, we get $K(f, g \frac{f_a}{g_a}) = K(g, f \frac{g_a}{f_a}) = K(f \frac{g_a}{f_a}, g)$ and $K(f_a, g_a) = K(g_a, f_a)$. Thus, proposition 7 implies that finding the argument of the maximum of $K(g_a, f_a)$ amounts to finding the argument of the maximum of $K(f_a, g_a)$. Consequently, the criteria of Huber's methodologies is $a \mapsto K(g_a, f_a)$. Second, our criteria is $a \mapsto K(g \frac{g_a}{f_a}, f)$ and property 4 implies $K(g, f \frac{g_a}{f_a}) = K(g \frac{f_a}{g_a}, f)$. Consequently, since [BROKEZ] takes into account the very form of the criteria, we are then in a position to compare Huber's methodologies with ours.*

To recapitulate, the choice of $r = \frac{f_a}{g_a}$ enables us to simultaneously solve the following four optimisation problems, for $a \in \mathbb{R}_*^d$:

First, find a such that $a = \arg \inf_{a \in \mathbb{R}_*^d} K(f \frac{g_a}{f_a}, g)$ - analytic approach -

Second, find a such that $a = \arg \inf_{a \in \mathbb{R}_*^d} K(f, g \frac{f_a}{g_a})$ - synthetic approach -

Third, find a such that $a = \arg \sup_{a \in \mathbb{R}_*^d} K(g_a, f_a)$ - to compare Huber's methods with ours -

Fourth, find a such that $a = \arg \inf_{a \in \mathbb{R}_*^d} K(g \frac{f_a}{g_a}, f)$ - our method.

4.2. On the sequence of the transformed densities $(g^{(j)})$

As already explained in the introduction section, the Mu Zhu article leads us to only consider Huber's synthetic approach.

4.2.1. Estimation of the first co-vector of f

Using the subsample X_1, X_2, \dots, X_n , see Appendix B, and following [BROKEZ], let us introduce the estimate of $K(g_a, f_{a,n})$, through $\hat{K}(g_a, f_{a,n}) = \int m(a, a, x) \left(\frac{f_{a,n}(a^\top x)}{f_n(x)} \right) d\mathbb{P}_n$

Proposition 8 *Let $\hat{a} := \arg \sup_{a \in \mathbb{R}_*^d} \hat{K}(g_a, f_{a,n})$.*

Then, \hat{a} is a strongly convergent estimate of a , as defined in proposition 7.

Finally, we define the following sequences $(\hat{a}_k)_{k \geq 1}$ and $(\hat{g}_n^{(k)})_{k \geq 1}$ - for any given n :

- \hat{a}_k is an estimate of a_k as defined in proposition 8 with $\hat{g}_n^{(k-1)}$ instead of g ,
- $\hat{g}_n^{(k)}$ is such that $\hat{g}_n^{(0)} = g$ and $\hat{g}_n^{(k)}(x) = \hat{g}_n^{(k-1)}(x) \frac{f_{\hat{a}_k, n}(\hat{a}_k^\top x)}{[\hat{g}_n^{(k-1)}]_{\hat{a}_k, n}(\hat{a}_k^\top x)}$, i.e. $\hat{g}_n^{(k)}(x) = g(x) \prod_{j=1}^k \frac{f_{\hat{a}_j, n}(\hat{a}_j^\top x)}{[\hat{g}_n^{(j-1)}]_{\hat{a}_j, n}(\hat{a}_j^\top x)}$.

4.2.2. Convergence study at the k^{th} step of the algorithm

Let $\hat{b}_n(a) = \arg \sup_{b \in \Theta} \mathbb{P}_n^a m(b, a)$, with $a \in \Theta$, and $\hat{\beta}_n = \arg \sup_{a \in \Theta} \sup_{b \in \Theta} \mathbb{P}_n^a m(b, a)$, then

Proposition 9 Both $\sup_{a \in \Theta} \|\hat{b}_n(a) - a_k\|$ and $\hat{\beta}_n$ converge toward a_k a.s.

Finally, the following theorem shows that $\hat{g}_n^{(k)}$ converges almost everywhere towards $g^{(k)}$:

Theorem 6 For any given k , it holds $\hat{g}_n^{(k)} \rightarrow_n g^{(k)}$ a.s.

4.2.3. Asymptotic inference at the k^{th} step of the algorithm

The following theorem shows that $\hat{g}_n^{(k)}$ converges towards $g^{(k)}$ at the rate $O_{\mathbf{P}}(m^{-\frac{1}{4+d}})$ in three different cases, namely for any given x , with the L^1 distance and with the Kullback-Leibler divergence:

Theorem 7 It holds $|\hat{g}_n^{(k)}(x) - g^{(k)}(x)| = O_{\mathbf{P}}(m^{-\frac{1}{4+d}})$, $\int |\hat{g}_n^{(k)}(x) - g^{(k)}(x)| dx = O_{\mathbf{P}}(m^{-\frac{1}{4+d}})$ and $|K(f, \hat{g}_n^{(k)}) - K(f, g^{(k)})| = O_{\mathbf{P}}(m^{-\frac{1}{4+d}})$.

The following theorem shows that the laws of Huber's estimators of a_k , namely $\hat{b}_n(a_k)$ and $\hat{\beta}_n$, converge towards a linear combination of Gaussian variables.

Theorem 8 It holds $\sqrt{n} \mathcal{D}(\hat{b}_n(a_k) - a_k) \xrightarrow{\mathcal{L}^{aw}} \mathcal{E} \cdot \mathcal{N}_d(0, \mathbf{P} \|\frac{\partial}{\partial b} m(a_k, a_k)\|^2) + \mathcal{F} \cdot \mathcal{N}_d(0, \mathbf{P} \|\frac{\partial}{\partial a} m(a_k, a_k)\|^2)$ and $\sqrt{n} \mathcal{D}(\hat{\beta}_n - a_k) \xrightarrow{\mathcal{L}^{aw}} \mathcal{G} \cdot \mathcal{N}_d(0, \mathbf{P} \|\frac{\partial}{\partial a} m(a_k, a_k)\|^2) + \mathcal{F} \cdot \mathcal{N}_d(0, \mathbf{P} \|\frac{\partial}{\partial b} m(a_k, a_k)\|^2)$ where $\mathcal{E} = \mathbf{P} \frac{\partial^2}{\partial a^2} m(a_k, a_k)$, $\mathcal{F} = \mathbf{P} \frac{\partial^2}{\partial a \partial b} m(a_k, a_k)$, $\mathcal{G} = \mathbf{P} \frac{\partial^2}{\partial b^2} m(a_k, a_k)$ and $\mathcal{D} = (\mathbf{P} \frac{\partial^2}{\partial b^2} m(a_k, a_k) \mathbf{P} \frac{\partial^2}{\partial a^2} m(a_k, a_k) - \mathbf{P} \frac{\partial^2}{\partial a \partial b} m(a_k, a_k) \mathbf{P} \frac{\partial^2}{\partial b \partial a} m(a_k, a_k)) > 0$.

4.3. A stopping rule for the procedure

We first give an estimate of f . Then, we provide a stopping rule for this identification procedure.

Remark 6 In the case where f is known, as explained in section 14 of [HUB85], the sequence $(K(g_{a_k}^{(k-1)}, f_{a_k}))_{k \geq 1}$ converges towards zero. Many authors have studied this hypothesis and its consequences. For example, Huber deducts that, if f can be deconvoluted with a Gaussian component, $(K(g_{a_k}^{(k-1)}, f_{a_k}))_{k \geq 1}$ converges toward 0. He then shows that $g^{(i)}$ uniformly converges in L^1 towards f - see propositions 14.2 and 14.3 page 461 of his article.

4.3.1. Estimation of f

The following lemma shows that $\lim_k K(\hat{g}_{a_k, n}^{(k)}, f_{a_k, n})$ converges towards zero as k goes to infinity and thereafter as n goes to infinity :

Lemma 2 We have $\lim_n \lim_k K(\hat{g}_{a_k, n}^{(k)}, f_{a_k, n}) = 0$, a.s.

Then, the following theorem enables us to provide simulations through an estimation of f

Theorem 9 We have $\lim_n \lim_k \hat{g}_n^{(k)} = f$, a.s.

4.3.2. Testing of the criteria

In this paragraph, through a test of Huber's criteria, namely $a \mapsto K(\hat{g}_{a,n}^{(k)}, f_{a,n})$, we will build a stopping rule for the procedure. First, the next theorem gives us the law of Huber's criteria.

Theorem 10 *For a fixed k , we have*

$$\sqrt{n}(\text{Var}_{\mathbf{P}}(m(\hat{b}_n(\hat{\beta}_n), \hat{\beta}_n)))^{-1/2}(\mathbb{P}_n m(\hat{b}_n(\hat{\beta}_n), \hat{\beta}_n) - \mathbb{P}_n m(a_k, a_k)) \xrightarrow{\mathcal{L}^{aw}} \mathcal{N}(0, I),$$

as n goes to infinity, where k represents the k^{th} step of the algorithm and I is the identity matrix in \mathbb{R}^d .

Note that k is fixed in theorem 10 since $\hat{\beta}_n = \arg \sup_{a \in \Theta} \sup_{b \in \Theta} \mathbb{P}_n^a m(b, a)$ where m is a known function of k - see section 4.1. Thus, in the case where $K(g_a^{(k)}, f_a) = 0$, we obtain

Corollary 3

We have $\sqrt{n}(\text{Var}_{\mathbf{P}}(m(\hat{b}_n(\hat{\beta}_n), \hat{\beta}_n)))^{-1/2}(\mathbb{P}_n m(\hat{b}_n(\hat{\beta}_n), \hat{\beta}_n)) \xrightarrow{\mathcal{L}^{aw}} \mathcal{N}(0, I)$.

Hence, we propose the test of the null hypothesis $(H_0) : K(g_{a_k}^{(k-1)}, f_{a_k}) = 0$ versus the alternative $(H_1) : K(g_{a_k}^{(k-1)}, f_{a_k}) \neq 0$. Based on this result, we stop the algorithm, then, defining a_k as the last vector generated from the Huber's algorithm, we derive from corollary 3, a α -level confidence ellipsoid around a_k , namely $\mathcal{E}'_k = \{b \in \mathbb{R}^d; \sqrt{n}(\text{Var}_{\mathbf{P}}(m(b, b)))^{-1/2} \mathbb{P}_n m(b, b) \leq q_{\alpha}^{\mathcal{N}(0,1)}\}$ where $q_{\alpha}^{\mathcal{N}(0,1)}$ is the quantile of a α -level reduced centered normal distribution and where \mathbb{P}_n is the empirical measure arising from a realization of the sequences (X_1, \dots, X_n) and (Y_1, \dots, Y_n) .

Consequently, the following corollary provides us with a confidence region for the above test:

Corollary 4 \mathcal{E}'_k *is a confidence region for the test of the null hypothesis (H_0) versus (H_1) .*

5. Simulations

We illustrate this section by detailing three simulations.

In each simulation, the program follows our algorithm and aims at creating a sequence of densities $(g^{(j)})$, $j = 1, \dots, k$, $k < d$, such that $g(0) = g$, $g^{(j)} = g^{(j-1)} f_{a_j} / [g^{(j-1)}]_{a_j}$ and $K(g^{(k)}, f) = 0$, where K is the Kullback-Leibler divergence and $a_j = \arg \inf_b K(g^{(j-1)} f_b / [g^{(j-1)}]_b, f)$, for all $j = 1, \dots, k$.

Then, in the first two simulations, the program follows Huber's method and generates a sequence of densities $(g^{(j)})$, $j = 1, \dots, k$, $k < d$, such that $g(0) = g$, $g^{(j)} = g^{(j-1)} f_{a_j} / [g^{(j-1)}]_{a_j}$ and $K(f, g^{(k)}) = 0$, where K is the Kullback-Leibler divergence and $a_j = \arg \sup_b K([g^{(j-1)}]_b, f_b)$, for all $j = 1, \dots, k$.

Finally, in the third example, we study the robustness of our method with four outliers.

Simulation 1

We are in dimension 3(=d). We consider a sample of 50(=n) values of a random variable X with density f defined by,

$$f(x) = \text{Normal}(x_1 + x_2). \text{Gumbel}(x_0 + x_2). \text{Gumbel}(x_0 + x_1),$$

where the Gumbel law parameters are $(-3, 4)$ and $(1, 1)$ and where the normal distribution parameters are $(-5, 2)$. We generate a Gaussian random variable Y with a density - that we will name g - which has the same mean and variance as f .

In the first part of the program, we theoretically obtain $k = 2$, $a_1 = (1, 0, 1)$ and $a_2 = (1, 1, 0)$ (or $a_2 = (1, 0, 1)$ and $a_1 = (1, 1, 0)$ which leads us to the same conclusion). To get this result, we perform the following test

$$(H_0) : (a_1, a_2) = ((1, 0, 1), (1, 1, 0)) \text{ versus } (H_1) : (a_1, a_2) \neq ((1, 0, 1), (1, 1, 0)).$$

Moreover, if i represents the last iteration of the algorithm, then

$$\sqrt{n}(\text{Var}_{\mathbf{P}}(M(c_n(\gamma_n), \gamma_n)))^{(-1/2)} \mathbb{P}_n M(c_n(\gamma_n), \gamma_n) \xrightarrow{\mathcal{L}^{aw}} \mathcal{N}(0, 1),$$

and then we estimate (a_1, a_2) with the following 0.9(= α) level confidence ellipsoid

$$\mathcal{E}_i = \{b \in \mathbb{R}^3; (\text{Var}_{\mathbf{P}}(M(b, b)))^{-1/2} \mathbb{P}_n M(b, b) \leq q_{\alpha}^{\mathcal{N}(0,1)} / \sqrt{n} \simeq \frac{0.2533}{7.0710678} = 0.03582203\}.$$

Indeed, if $i = 1$ represents the last iteration of the algorithm, then $a_1 \in \mathcal{E}_1$, and if $i = 2$ represents the last iteration of the algorithm, then $a_2 \in \mathcal{E}_2$, and so on, if i represents the last iteration of the algorithm, then $a_i \in \mathcal{E}_i$.

Now, if we follow Huber's method, we also theoretically obtain $k = 2$, $a_1 = (1, 0, 1)$ and $a_2 = (1, 1, 0)$ (or $a_2 = (1, 0, 1)$ and $a_1 = (1, 1, 0)$ which leads us to the same conclusion). To get this result, we perform the following test:

$$(H_0) : (a_1, a_2) = ((1, 0, 1), (1, 1, 0)) \text{ versus } (H_1) : (a_1, a_2) \neq ((1, 0, 1), (1, 1, 0)).$$

Similarly as above, the fact that, if i represents the last iteration of the algorithm, then

$\sqrt{n}(\text{Var}_{\mathbf{P}}(m(b_n(\beta_n), \beta_n)))^{(-1/2)} \mathbb{P}_n m(b_n(\beta_n), \beta_n) \xrightarrow{\mathcal{L}^{aw}} \mathcal{N}(0, 1)$, enables us to estimate our sequence of (a_i) , reduced to (a_1, a_2) , through the following 0.9(= α) level confidence ellipsoid

$$\mathcal{E}'_i = \{b \in \mathbb{R}^3; (\text{Var}_{\mathbf{P}}(m(b, b)))^{-1/2} \mathbb{P}_n m(b, b) \leq q_{\alpha}^{\mathcal{N}(0,1)} / \sqrt{n} \simeq 0.03582203\}.$$

Finally, we obtain

Table 1: Simulation 1 : Numerical results of the optimisation.

	Our Algorithm	Huber's Algorithm
Projection Study 0 :	minimum : 0.317505	maximum : 0.715135
	at point : (1.0,1.0,0)	at point : (1.0,1.0,0)
	P-Value : 0.99851	P-Value : 0.999839
Test :	$H_0 : a_1 \in \mathcal{E}_1$: False	$H_0 : a_1 \in \mathcal{E}'_1$: False
Projection Study 1 :	minimum : 0.0266514	maximum : 0.007277
	at point : (1.0,0,1.0)	at point : (1.0,0,1.0)
	P-Value : 0.998852	P-Value : 0.999835
Test :	$H_0 : a_2 \in \mathcal{E}_2$: True	$H_0 : a_2 \in \mathcal{E}'_2$: True
K(Estimate $g_m^{(2)}, g^{(2)}$)	0.444388	0.794124

Therefore, we conclude that $f = g^{(2)}$.

Simulation 2

We are in dimension 10(=d). We consider a sample of 50(=n) values of a random variable X with density f defined by,

$$f(x) = \text{Gumbel}(x_0).Normal(x_1, \dots, x_9),$$

where the Gumbel law parameters are -5 and 1 and where the normal distribution is reduced and centered.

Our reasoning is the same as in Example 1. In the first part of the program, we theoretically obtain $k = 1$ and $a_1 = (1, 0, \dots, 0)$. To get this result, we perform the following test

$$(H_0) : a_1 = (1, 0, \dots, 0) \text{ versus } (H_1) : a_1 \neq (1, 0, \dots, 0).$$

We estimate a_1 by the following 0.9(= α) level confidence ellipsoid

$$\mathcal{E}_i = \{b \in \mathbb{R}^2; (\text{Var}_{\mathbf{P}}(M(b, b)))^{-1/2} \mathbb{P}_n M(b, b) \leq q_{\alpha}^{\mathcal{N}(0,1)} / \sqrt{n} \simeq 0.03582203\}.$$

Now, if we follow Huber's method, we also theoretically obtain $k = 1$ and $a_1 = (1, 0, \dots, 0)$. To get this result, we perform the following test

$(H_0) : a_1 = (1, 0, \dots, 0)$ versus $(H_1) : a_1 \neq (1, 0, \dots, 0)$.

Hence, using the same reasoning as in Example 1, we estimate a_1 through the following 0.9 ($=\alpha$) level confidence ellipsoid

$$\mathcal{E}'_i = \{b \in \mathbb{R}^2; (Var_{\mathbf{P}}(m(b, b)))^{-1/2} \mathbb{P}_n m(b, b) \leq q_{\alpha}^{\mathcal{N}(0,1)} / \sqrt{n} \simeq 0.03582203\}.$$

And, we obtain

Table 2: Simulation 2 : Numerical results of the optimisation.

	Our Algorithm	Huber's Algorithm
Projection Study 0:	minimum : 0.00263554	maximum : 0.00376235
	at point : (1.0001,	at point : (0.9902,
	0.0040338, 0.098606, 0.115214,	0.0946806, 0.161447, 0.0090245,
	0.067628, 0.16229, 0.00549203,	0.147804, 0.180259, 0.0975065,
	0.014319, 0.149339, 0.0578906)	0.101044, 0.190976, 0.155706)
	P-Value : 0.828683	P-Value : 0.807121
Test :	$H_0 : a_1 \in \mathcal{E}_1 : \text{True}$	$H_0 : a_1 \in \mathcal{E}'_1 : \text{True}$
K(Estimate $g_m^{(1)}, g^{(1)}$)	2.44546	2.32331

Therefore, we conclude that $f = g^{(1)}$.

Simulation 3

We are in dimension 20(=d). We first generate a sample with 100(=n) observations, namely four outliers $x = (2, 0, \dots, 0)$ and 96 values of a random variable X with a density f defined by

$$f(x) = \text{Gumbel}(x_0).Normal(x_1, \dots, x_{19})$$

where the Gumbel law parameters are -5 and 1 and where the normal distribution is reduced and centered. Our reasoning is the same as in Simulation 1.

We theoretically obtain $k = 1$ and $a_1 = (1, 0, \dots, 0)$. To get this result, we perform the following test

$$(H_0) : a_1 = (1, 0, \dots, 0) \text{ versus } (H_1) : a_1 \neq (1, 0, \dots, 0)$$

We estimate a_1 by the following 0.9($=\alpha$) level confidence ellipsoid

$$\mathcal{E}_i = \{b \in \mathbb{R}^2; (Var_{\mathbf{P}}(M(b, b)))^{-1/2} \mathbb{P}_n M(b, b) \leq q_{\alpha}^{\mathcal{N}(0,1)} / \sqrt{n} \simeq 0.02533\}$$

And, we obtain

Table 3: Simulation 3: Numerical results of the optimisation.

Our Algorithm	
Projection Study 0	minimum : 0.024110
	at point : (0.8221, 0.0901, 0.0892, -0.2020, 0.0039, 0.1001,
	0.0391, 0.08001, 0.07633, -0.0437, 0.12093, 0.09834, 0.1045,
	0.0874, -0.02349, 0.03001, 0.12543, 0.09435, 0.0587, -0.0055)
	P-Value : 0.77004
Test :	$H_0 : a_1 \in \mathcal{E}_1 : \text{True}$
K(Estimate $g_m^{(1)}, g^{(1)}$)	2.677015

Therefore, we conclude that $f = g^{(1)}$.

Critics of the simulations

As customary in simulation studies, as approximations accumulate, results depend on the power of the calculators used as well as on the available memory. Moreover, in order to implement our optimisation in \mathbb{R}^d of the relative entropy, we choose to apply the simulated annealing method.

Thus, in the case where f is unknown, we will never have the certainty to have reached the desired minimum or maximum of the Kullback-Leibler divergence. Indeed, this probabilistic metaheuristic only converges, and the probability to reach the minimum or the maximum only tends towards 1, when the number of random jumps tends in theory towards infinity.

We also note that no theory on the optimal number of jumps to implement does exist, as this number depends on the specificities of each particular problem.

Finally, we choose the $50^{-\frac{4}{4+d}}$ (resp. $100^{-\frac{4}{4+d}}$) for the AMISE of the simulations 1 and 2 (resp. 3). This choice leads us to simulate 50 (resp. 100) random variables, see [SCOTT92] page 151, none of which have been discarded to obtain the truncated sample.

Conclusion

Characteristic structures as well as one-dimensional projections and their associated distributions in multivariate datasets can be evidenced through Projection Pursuit.

The present article demonstrates that our Kullback-Leibler divergence minimisation method constitutes a good alternative to Huber's relative entropy maximization approach, see [HUB85]. Indeed, the convergence results as well as the simulations we carried out clearly evidences the robustness of our methodology.

A. Reminders

A.1. The relative entropy (or Kullback-Leibler divergence)

We call h_a the density of $a^\top Z$ if h is the density of Z , and K the relative entropy or Kullback-Leibler divergence. The function K is defined by - considering P and Q , two probabilities:

$$K(Q, P) = \int \varphi\left(\frac{\partial Q}{\partial P}\right) dP \text{ if } P \ll Q \text{ and}$$

$$K(Q, P) = +\infty \text{ otherwise,}$$

where $\varphi : x \mapsto x \ln(x) - x + 1$ is strictly convex.

Let us present some well-known properties of the Kullback-Leibler divergence.

Property 2 *We have $K(P, Q) = 0 \Leftrightarrow P = Q$.*

Property 3 *The divergence function $Q \mapsto K(Q, P)$ is convex, lower semi-continuous (l.s.c.) - for the topology that makes all the applications of the form $Q \mapsto \int f dQ$ continuous where f is bounded and continuous - as well as l.s.c. for the topology of the uniform convergence.*

Property 4 (corollary (1.29), page 19 of [LIVAJ]) *If $T : (X, A) \rightarrow (Y, B)$ is measurable and if $K(P, Q) < \infty$, then $K(P, Q) \geq K(PT^{-1}, QT^{-1})$, with equality being reached when T is surjective for (P, Q) .*

Theorem 11 (theorem III.4 of [AZE97]) *Let $f : I \rightarrow \mathbb{R}$ be a convex function. Then f is a Lipschitz function in all compact intervals $[a, b] \subset \text{int}\{I\}$. In particular, f is continuous on $\text{int}\{I\}$.*

A.2. Useful lemmas

Lemma 3 Let f be a density in \mathbb{R}^d bounded and positive. Then, any projection density of f - that we will name f_a , with $a \in \mathbb{R}_*^d$ - is also bounded and positive in \mathbb{R} .

Lemma 4 Let f be a density in \mathbb{R}^d bounded and positive. Then any density $f(./a^\top x)$, for any $a \in \mathbb{R}_*^d$, is also bounded and positive.

Lemma 5 If f and g are positive and bounded densities, then $g^{(k)}$ is positive and bounded.

Lemma 6 Let f be an absolutely continuous density, then, for all sequences (a_n) tending to a in \mathbb{R}_*^d , the sequence f_{a_n} uniformly converges towards f_a .

Proof :

For all a in \mathbb{R}_*^d , let F_a be the cumulative distribution function of $a^\top X$ and ψ_a be a complex function defined by $\psi_a(u, v) = F_a(\mathcal{R}e(u + iv)) + iF_a(\mathcal{R}e(v + iu))$, for all u and v in \mathbb{R} .

First, the function $\psi_a(u, v)$ is an analytic function, because $x \mapsto f_a(a^\top x)$ is continuous and as a result of the corollary of Dini's second theorem - according to which "A sequence of cumulative distribution functions which pointwise converges on \mathbb{R} towards a continuous cumulative distribution function F on \mathbb{R} , uniformly converges towards F on \mathbb{R} " - we deduct that, for all sequences (a_n) converging towards a , ψ_{a_n} uniformly converges towards ψ_a . Finally, the Weierstrass theorem, (see proposal (10.1) page 220 of [DI80]), implies that all sequences $\psi'_{a,n}$ uniformly converge towards ψ'_a , for all a_n tending to a . We can therefore conclude. \square

Lemma 7 The set Γ_c is closed in L^1 for the topology of the uniform convergence.

Lemma 8 For all $c > 0$, we have $\Gamma_c \subset \overline{B}_{L^1}(f, c)$, where $B_{L^1}(f, c) = \{p \in L^1; \|f - p\|_1 \leq c\}$.

Lemma 9 G is closed in L^1 for the topology of the uniform convergence.

Lemma 10 Let H be an integrable function and let $C = \int H d\mathbf{P}$ and $C_n = \int H d\mathbb{P}_n$, then, $C_n - C = O_{\mathbf{P}}(\frac{1}{\sqrt{n}})$.

B. Study of the sample

Let X_1, X_2, \dots, X_m be a sequence of independent random vectors with the same density f . Let Y_1, Y_2, \dots, Y_m be a sequence of independent random vectors with the same density g . Then, the kernel estimators f_m , and $f_{a,m}$ of f and f_a , for all $a \in \mathbb{R}_*^d$, almost surely and uniformly converge since we assume that the bandwidth h_m of these estimators meets the following conditions (see [BOLE]):

(*Hyp*): $h_m \searrow_m 0$, $mh_m \nearrow_m \infty$, $mh_m/L(h_m^{-1}) \rightarrow_m \infty$ and $L(h_m^{-1})/LLm \rightarrow_m \infty$, with $L(u) = \ln(u \vee e)$.

Let us consider $A_0(m, a) = \frac{1}{m} \sum_{i=1}^m \ln \left\{ \frac{g_a(a^\top Y_i)}{f_{a,m}(a^\top Y_i)} \right\} \frac{g_a(a^\top Y_i)}{g(Y_i)}$, $A'_0(m, a) = \frac{1}{m} \sum_{i=1}^m \left(\frac{g_a(a^\top X_i)}{f_{a,m}(a^\top X_i)} - 1 \right) \frac{f_{a,m}(a^\top X_i)}{f_m(X_i)}$, $B_0(m, a) = \frac{1}{m} \sum_{i=1}^m \ln \left\{ \frac{f_{a,m}(a^\top Y_i)}{g_a(a^\top Y_i)} \frac{g(Y_i)}{f_m(Y_i)} \right\} \frac{f_{a,m}(a^\top Y_i)}{g_a(a^\top Y_i)}$, $B'_0(m, a) = \frac{1}{m} \sum_{i=1}^m \left(1 - \left\{ \frac{f_{a,m}(a^\top X_i)}{g_a(a^\top X_i)} \frac{g(X_i)}{f_m(X_i)} \right\} \right)$.

Our goal is to estimate the maximum of $K(g_a, f_a)$ and the minimum of $K(g \frac{f_a}{g_a}, f)$.

To achieve this, it is necessary for us to truncate X_1, X_2, \dots, X_m and Y_1, Y_2, \dots, Y_m :

Let us consider now a sequence θ_m such that $\theta_m \rightarrow 0$, and $y_m/\theta_m^2 \rightarrow 0$, where y_m is defined through lemma 13 with $y_m = O_{\mathbf{P}}(m^{-\frac{2}{4+d}})$. We will generate f_m and $f_{b,m}$ from the starting sample and we select the X_i and the Y_i vectors such that $f_m(X_i) \geq \theta_m$ and $g(Y_i) \geq \theta_m$, for all i and for all $b \in \mathbb{R}_*^d$ - for Huber's algorithm - and such that $f_m(X_i) \geq \theta_m$ and $g_b(b^\top Y_i) \geq \theta_m$, for all i and for all $b \in \mathbb{R}_*^d$ - for our algorithm. The vectors meeting these conditions will be called X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_n . Consequently, the next proposition provides us with the condition required to obtain our estimates

Proposition 10 *Using the notations introduced in [BROKEZ] and in sections 3.1.1. and 4.1., it holds*

$$\sup_{a \in \mathbb{R}_*^d} |(A_0(n, a) - A'_0(n, a)) - K(g_a, f_a)| \rightarrow 0 \text{ a.s.}, \quad (6)$$

$$\sup_{a \in \mathbb{R}_*^d} |(B_0(n, a) - B'_0(n, a)) - K(g \frac{f_a}{g_a}, f)| \rightarrow 0 \text{ a.s.} \quad (7)$$

Remark 7 *We can take for θ_m the expression $m^{-\nu}$, with $0 < \nu < \frac{1}{4+d}$. Moreover, to estimate a_k , $k \geq 2$, we use the same procedure than the one we followed in order to find a_1 with $g_n^{(k-1)}$ instead of g - since $g^{(k-1)}$ is unknown in this case.*

C. Case study : f is known

In this Appendix, we study the case when f and g are known.

C.1. Convergence study at the k^{th} step of the algorithm:

In this paragraph, when k is less than or equal to d , we show that the sequence $(\check{a}_k)_n$ converges towards a_k and that the sequence $(\check{g}^{(k)})_n$ converges towards $g^{(k)}$.

Both $\check{\gamma}_n$ and $\check{c}_n(a)$ are M-estimators and estimate a_k - see [BROKEZ]. We state

Proposition 11 *Assuming $(H'1)$ to $(H'3)$ hold. Both $\sup_{a \in \Theta} \|\check{c}_n(a) - a_k\|$ and $\check{\gamma}_n$ tends to a_k a.s.*

Finally, the following theorem shows us that $\check{g}^{(k)}$ converges uniformly almost everywhere towards $g^{(k)}$, for any $k = 1..d$.

Theorem 12 *Assuming $(H'1)$ to $(H'3)$ hold. Then, $\check{g}^{(k)} \rightarrow_n g^{(k)}$ a.s. and uniformly a.e.*

C.2. Asymptotic Inference at the k^{th} step of the algorithm

The following theorem shows that $\check{g}^{(k)}$ converges at the rate $O_{\mathbf{P}}(n^{-1/2})$ in three different cases, namely for any given x , with the L^1 distance and with the Kullback-Leibler divergence:

Theorem 13 *Assuming $(H'0)$ to $(H'3)$ hold, for any $k = 1, \dots, d$ and any $x \in \mathbb{R}^d$, we have*

$$|\check{g}^{(k)}(x) - g^{(k)}(x)| = O_{\mathbf{P}}(n^{-1/2}), \quad (8)$$

$$\int |\check{g}^{(k)}(x) - g^{(k)}(x)| dx = O_{\mathbf{P}}(n^{-1/2}), \quad (9)$$

$$|K(\check{g}^{(k)}, f) - K(g^{(k)}, f)| = O_{\mathbf{P}}(n^{-1/2}). \quad (10)$$

The following theorem shows that the laws of our estimators of a_k , namely $\check{c}_n(a_k)$ and $\check{\gamma}_n$, converge towards a linear combination of Gaussian variables.

Theorem 14 *Assuming that conditions $(H'1)$ to $(H'6)$ hold, then*

$\sqrt{n}\mathcal{A}(\check{c}_n(a_k) - a_k) \xrightarrow{\mathcal{L}^{aw}} \mathcal{B}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}M(a_k, a_k)\|^2) + \mathcal{C}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}M(a_k, a_k)\|^2)$ and

$\sqrt{n}\mathcal{A}(\check{\gamma}_n - a_k) \xrightarrow{\mathcal{L}^{aw}} \mathcal{C}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}M(a_k, a_k)\|^2) + \mathcal{C}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}M(a_k, a_k)\|^2)$

where $\mathcal{A} = (\mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k, a_k)(\mathbf{P}\frac{\partial^2}{\partial a_i \partial a_j}M(a_k, a_k) + \mathbf{P}\frac{\partial^2}{\partial a_i \partial b_j}M(a_k, a_k)))$,

$\mathcal{C} = \mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k, a_k)$ and $\mathcal{B} = \mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k, a_k) + \mathbf{P}\frac{\partial^2}{\partial a_i \partial a_j}M(a_k, a_k) + \mathbf{P}\frac{\partial^2}{\partial a_i \partial b_j}M(a_k, a_k)$.

C.3.A stopping rule for the procedure

We now assume that the algorithm does not stop after d iterations. We then remark that, it still holds - for any $i > d$:

- $g^{(i)}(x) = g(x) \prod_{k=1}^i \frac{f_{a_k}(a_k^\top x)}{[g_n^{(k-1)}]_{a_k}(a_k^\top x)}$, with $g^{(0)} = g$.
- $K(g^{(0)}, f) \geq K(g^{(1)}, f) \geq K(g^{(2)}, f) \dots \geq 0$.
- Theorems 12, 13 and 14.

Moreover, through remark 5 page 10 and as explained in section 14 of [HUB85], the sequence $(K(g^{(k-1)} \frac{f_{a_k}}{g_{a_k}^{(k-1)}}), f)_{k \geq 1}$ converges towards zero. Then, in this paragraph, we show that $g^{(i)}$ converges towards f in i . Finally, we provide a stopping rule for this identification procedure.

C.3.1. Representation of f

Under $(H'0)$, the following proposition shows us that the probability measure with density $g^{(k)}$ converges towards the probability measure with density f :

Proposition 12 *We have $\lim_k g^{(k)} = f$ a.s.*

C.3.2. Testing of the criteria

Through a test of the criteria, namely $a \mapsto K(g^{(k-1)} \frac{f_a}{g_a^{(k-1)}}), f)$, we build a stopping rule for this procedure. First, the next theorem enables us to derive the law of the criteria.

Theorem 15 *Assuming that $(H'1)$ to $(H'3)$, $(H'6)$ and $(H'8)$ hold. Then,*

$$\sqrt{n}(\text{Var}_{\mathbf{P}}(M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n)))^{-1/2}(\mathbb{P}_n M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n) - \mathbb{P}_n M(a_k, a_k)) \xrightarrow{\mathcal{L}^{aw}} \mathcal{N}(0, I),$$

where k represents the k^{th} step of the algorithm and with I being the identity matrix in \mathbb{R}^d .

Note that k is fixed in theorem 15 since $\check{\gamma}_n = \arg \inf_{a \in \Theta} \sup_{c \in \Theta} \mathbb{P}_n M(c, a)$ where M is a known function of k - see section 3.1.1. Thus, in the case where $K(g^{(k-1)} \frac{f_{a_k}}{g_{a_k}^{(k-1)}}), f) = 0$, we obtain

Corollary 5 *Assuming that $(H'1)$ to $(H'3)$, $(H'6)$, $(H'7)$ and $(H'8)$ hold. Then,*

$$\sqrt{n}(\text{Var}_{\mathbf{P}}(M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n)))^{-1/2}(\mathbb{P}_n M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n)) \xrightarrow{\mathcal{L}^{aw}} \mathcal{N}(0, I).$$

Hence, we propose the test of the null hypothesis $(H_0) : K(g^{(k-1)} \frac{f_{a_k}}{g_{a_k}^{(k-1)}}), f) = 0$ versus $(H_1) : K(g^{(k-1)} \frac{f_{a_k}}{g_{a_k}^{(k-1)}}), f) \neq 0$. Based on this result, we stop the algorithm, then, defining a_k as the last vector generated, we derive from corollary 5 a α -level confidence ellipsoid around a_k , namely

$$\mathcal{E}_k = \{b \in \mathbb{R}^d; \sqrt{n}(\text{Var}_{\mathbf{P}}(M(b, b)))^{-1/2} \mathbb{P}_n M(b, b) \leq q_\alpha^{\mathcal{N}(0,1)}\},$$

where $q_\alpha^{\mathcal{N}(0,1)}$ is the quantile of a α -level reduced centered normal distribution.

Consequently, the following corollary provides us with a confidence region for the above test:

Corollary 6 \mathcal{E}_k is a confidence region for the test of the null hypothesis (H_0) versus (H_1) .

D. Hypotheses' discussion

D.1. Discussion on $(H'2)$.

We verify this hypothesis in the case where :

- a_1 is the unique element of \mathbb{R}_*^d such that $f(./a_1^\top x) = g(./a_1^\top x)$, i.e. $K(g(./a_1^\top x) f_{a_1}(a_1^\top x), f) = 0, (1)$
- f and g are bounded and positive, (2)
- there exists a neighborhood V of a_k such that, for all b in V and for all positive real A , there exists $\mathcal{S} > 0$ such that $g(./b^\top x) \leq \mathcal{S}.f(./b^\top x)$ with $\|x\| > A$ (3).

We remark that we obtain the same proof with f , $g^{(k-1)}$ and a_k .

First, (1) implies that $g_{g_{a_1}}^{f_{a_1}} = f$. Hence, $0 > \int \ln(\frac{g}{f} \frac{f_c}{g_c}) g_{g_{a_1}}^{f_{a_1}} dx = -K(g_{g_c}^{f_c}, f) > -K(g, f)$ as a result of the very construction of $g_{g_c}^{f_c}$. Besides, (2) and (3) imply that there exists a neighborhood V of a_k such that, for all c in V , there exists $\mathcal{S} > 0$ such that, for all x in \mathbb{R}^d , $g(\cdot/c'x) \leq \mathcal{S} \cdot f(\cdot/c'x)$.

Consequently, we get $|M(c, a_1, x)| \leq |-K(g, f)| + |- (\frac{g(\cdot/c'x)}{f(\cdot/c'x)} - 1)| \leq K(g, f) + \mathcal{S} + 1$.

Finally, we infer the existence a neighborhood V of a_k such that, for all c in V ,

$$|M(c, a_k, x)| \leq H(x) = K(g, f) + \mathcal{S} + 1 \text{ (P - a.s.) with } \mathbf{P}H < \infty.$$

D.2. Discussion on (H'3).

We verify this hypothesis in the case where a_1 is the unique element of \mathbb{R}_*^d such that $f(\cdot/a_1^\top x) = g(\cdot/a_1^\top x)$, i.e. $K(g(\cdot/a_1^\top x)f_{a_1}(a_1^\top x), f) = 0$ - we obtain the same proof with f , $g^{(k-1)}$ and a_k .

Preliminary (A): Shows that $A = \{(c, x) \in \mathbb{R}_^d \setminus \{a_1\} \times \mathbb{R}^d; \frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} > \frac{f_c(c^\top x)}{g_c(c^\top x)} \text{ and } g(x) \frac{f_c(c^\top x)}{g_c(c^\top x)} > f(x)\} = \emptyset$ through a reductio ad absurdum, i.e. if we assume $A \neq \emptyset$.*

Thus, we have $f(x) = f(\cdot/a_1^\top x)f_{a_1}(a_1^\top x) = g(\cdot/a_1^\top x)f_{a_1}(a_1^\top x) > g(\cdot/c^\top x)f_c(c^\top x) > f$, since $\frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} \geq \frac{f_c(c^\top x)}{g_c(c^\top x)}$ implies $g(\cdot/a_1^\top x)f_{a_1}(a_1^\top x) = g(x) \frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} \geq g(x) \frac{f_c(c^\top x)}{g_c(c^\top x)} = g(\cdot/c^\top x)f_c(c^\top x)$, i.e. $f > f$. We can therefore conclude.

Preliminary (B): Shows that $B = \{(c, x) \in \mathbb{R}_^d \setminus \{a_1\} \times \mathbb{R}^d; \frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} < \frac{f_c(c^\top x)}{g_c(c^\top x)} \text{ and } g(x) \frac{f_c(c^\top x)}{g_c(c^\top x)} < f(x)\} = \emptyset$ through a reductio ad absurdum, i.e. if we assume $B \neq \emptyset$.*

Thus, we have $f(x) = f(\cdot/a_1^\top x)f_{a_1}(a_1^\top x) = g(\cdot/a_1^\top x)f_{a_1}(a_1^\top x) < g(\cdot/c^\top x)f_c(c^\top x) < f$.

We can thus conclude as above.

Let us now prove (H'3). We have $PM(c, a_1) - PM(c, a) = \int \ln(\frac{g(x)f_c(c^\top x)}{g_c(c^\top x)f(x)}) \{ \frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} - \frac{f_c(c^\top x)}{g_c(c^\top x)} \} g(x) dx$.

Moreover, the logarithm \ln is negative on $\{x \in \mathbb{R}_*^d; \frac{g(x)f_c(c^\top x)}{g_c(c^\top x)f(x)} < 1\}$ and is positive on $\{x \in \mathbb{R}_*^d; \frac{g(x)f_c(c^\top x)}{g_c(c^\top x)f(x)} \geq 1\}$. Thus, the preliminary studies (A) and (B) show that $\ln(\frac{g(x)f_c(c^\top x)}{g_c(c^\top x)f(x)})$ and $\{ \frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} - \frac{f_c(c^\top x)}{g_c(c^\top x)} \}$ always present a negative product. We can thus conclude, since $(c, a) \mapsto PM(c, a_1) - PM(c, a)$ is not null for all c and for all $a \neq a_1$. \square

E. Proofs

Remark 8 1/ (H'0) - according to which f and g are assumed to be positive and bounded - through lemma 5 (see page 16) implies that $\check{g}_n^{(k)}$ and $\hat{g}_n^{(k)}$ are positive and bounded.

2/ Remark 4 implies that f_n , g_n , $\check{g}_n^{(k)}$ and $\hat{g}_n^{(k)}$ are positive and bounded since we consider a Gaussian kernel.

Proof of propositions 5 and 6. Let us first study proposition 6.

Without loss of generality, we prove this proposition with x_1 in lieu of $a^\top X$.

We define $g^* = gr$. We remark that g and g^* present the same density conditionally to x_1 . Indeed, $g_1^*(x_1) = \int g^*(x) dx_2 \dots dx_d = \int r(x_1) g(x) dx_2 \dots dx_d = r(x_1) \int g(x) dx_2 \dots dx_d = r(x_1) g_1(x_1)$.

Thus, we can prove this proposition. We have $g(\cdot|x_1) = \frac{g(x_1, \dots, x_n)}{g_1(x_1)}$ and $g_1(x_1)r(x_1)$ is the marginal density of g^* . Hence, g^* is a density since g^* is positive and since

$\int g^* dx = \int g_1(x_1) r(x_1) g(\cdot | x_1) dx = \int g_1(x_1) \frac{f_1(x_1)}{g_1(x_1)} (\int g(\cdot | x_1) dx_2 \dots dx_d) dx_1 = \int f_1(x_1) dx_1 = 1$. Moreover,

$$K(f, g^*) = \int f \{ \ln(f) - \ln(g^*) \} dx, \quad (11)$$

$$\begin{aligned} &= \int f \{ \ln(f(\cdot | x_1)) - \ln(g^*(\cdot | x_1)) + \ln(f_1(x_1)) - \ln(g_1(x_1)r(x_1)) \} dx, \\ &= \int f \{ \ln(f(\cdot | x_1)) - \ln(g(\cdot | x_1)) + \ln(f_1(x_1)) - \ln(g_1(x_1)r(x_1)) \} dx, \end{aligned} \quad (12)$$

as $g^*(\cdot | x_1) = g(\cdot | x_1)$. Since the minimum of this last equation (12) is reached through the minimization of $\int f \{ \ln(f_1(x_1)) - \ln(g_1(x_1)r(x_1)) \} dx = K(f_1, g_1 r)$, then property 2 necessarily implies that $f_1 = g_1 r$, hence $r = f_1/g_1$. Finally, we have $K(f, g) - K(f, g^*) = \int f \{ \ln(f_1(x_1)) - \ln(g_1(x_1)) \} dx = K(f_1, g_1)$, which completes the demonstration of proposition 6.

Similarly, if we replace $f^* = f r^{-1}$ with f and g with g^* , we obtain the proof of proposition 5. \square

Proof of propositions 2 and 7. The proof of proposition 2 (resp. 7) is very similar to the one for proposition 6, save for the fact we now base our reasoning at row 11 on $K(g^*, f) = \int g^* \{ \ln(f) - \ln(g^*) \} dx$ (resp. $\int g \{ \ln(g^*) - \ln(f) \} dx$) instead of $K(f, g^*) = \int f \{ \ln(f) - \ln(g^*) \} dx$. \square

Proof of lemma 11.

Lemma 11 *If the family $(a_i)_{i=1\dots d}$ is a basis of \mathbb{R}^d then*

$$g(\cdot / a_1^\top x, \dots, a_j^\top x) = n(a_{j+1}^\top x, \dots, a_d^\top x) = f(\cdot / a_1^\top x, \dots, a_j^\top x).$$

Putting $A = (a_1, \dots, a_d)$, let us determine f in the A basis. Let us first study the function defined by $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $x \mapsto (a_1^\top x, \dots, a_d^\top x)$. We can immediately say that ψ is continuous and since A is a basis, its bijectivity is obvious. Moreover, let us study its Jacobian. By definition, it is $J_\psi(x_1, \dots, x_d) = |(\frac{\partial \psi_i}{\partial x_j})_{1 \leq i, j \leq d}| = |(a_{i,j})_{1 \leq i, j \leq d}| = |A| \neq 0$ since A is a basis. We can therefore infer for any x in \mathbb{R}^d , there exists a unique y in \mathbb{R}^d such that $f(x) = |A|^{-1} \Psi(y)$, i.e. Ψ (resp. y) is the expression of f (resp of x) in basis A , namely $\Psi(y) = \tilde{n}(y_{j+1}, \dots, y_d) \tilde{h}(y_1, \dots, y_j)$, with \tilde{n} and \tilde{h} being the expressions of n and h in the A basis. Consequently, our results in the case where the family $\{a_j\}_{1 \leq j \leq d}$ is the canonical basis of \mathbb{R}^d , still hold for Ψ in the A basis - see section 2.1.2. And then, if \tilde{g} is the expression of g in the A basis, we have $\tilde{g}(\cdot / y_1, \dots, y_j) = \tilde{n}(y_{j+1}, \dots, y_d) = \Psi(\cdot / y_1, \dots, y_j)$, i.e. $g(\cdot / a_1^\top x, \dots, a_j^\top x) = n(a_{j+1}^\top x, \dots, a_d^\top x) = f(\cdot / a_1^\top x, \dots, a_j^\top x)$. \square

Proof of lemma 12.

Lemma 12 $\inf_{a \in \mathbb{R}_*^d} K(g \frac{f_a}{g_a}, f)$ is reached.

Indeed, let G be $\{g \frac{f_a}{g_a}; a \in \mathbb{R}_*^d\}$ and Γ_c be $\Gamma_c = \{p; K(p, f) \leq c\}$ for all $c > 0$. From lemmas 7, 8 and 9 (see page 16), we get $\Gamma_c \cap G$ is a compact for the topology of the uniform convergence, if $\Gamma_c \cap G$ is not empty. Hence, and since property 3 (see page 15) implies that $Q \mapsto K(Q, P)$ is lower semi-continuous in L^1 for the topology of the uniform convergence, then the infimum is reached in L^1 . (Taking for example $c = K(g, f)$, Ω is necessarily not empty because we always have $K(g \frac{f_a}{g_a}, f) \leq K(g, f)$). \square

Proof of lemma 13.

Lemma 13 *For any continuous density f , we have $y_m = |f_m(x) - f(x)| = O_{\mathbf{P}}(m^{-\frac{2}{4+d}})$.*

Defining $b_m(x)$ as $b_m(x) = |E(f_m(x)) - f(x)|$, we have $y_m \leq |f_m(x) - E(f_m(x))| + b_m(x)$. Moreover, from page 150 of [SCOTT92], we derive that $b_m(x) = O_{\mathbf{P}}(\sum_{j=1}^d h_j^2)$ where $h_j = O_{\mathbf{P}}(m^{-\frac{1}{4+d}})$. Then, we

infer $b_m(x) = O_{\mathbf{P}}(m^{-\frac{2}{4+d}})$. Finally, since the central limit theorem rate is $O_{\mathbf{P}}(m^{-\frac{1}{2}})$, we then obtain that $y_m \leq O_{\mathbf{P}}(m^{-\frac{1}{2}}) + O_{\mathbf{P}}(m^{-\frac{2}{4+d}}) = O_{\mathbf{P}}(m^{-\frac{2}{4+d}})$. \square

Proof of proposition 10. We prove this proposition for $k \geq 2$, i.e. in the case where $g^{(k-1)}$ is not known. The initial case using the known density $g^{(0)} = g$, will be an immediate consequence from the above. Moreover, going forward, to be more legible, we will use g (resp. g_n) in lieu of $g^{(k-1)}$ (resp. $g_n^{(k-1)}$). We can therefore remark that we have $f(X_i) \geq \theta_n - y_n$, $g(Y_i) \geq \theta_n - y_n$ and $g_b(b^\top Y_i) \geq \theta_n - y_n$, for all i and for all $b \in \mathbb{R}_*^d$, thanks to the uniform convergence of the kernel estimators. Indeed, we have $f(X_i) = f(X_i) - f_n(X_i) + f_n(X_i) \geq -y_n + f_n(X_i)$, by definition of y_n , and then $f(X_i) \geq -y_n + \theta_n$, by hypothesis on $f_n(X_i)$. This is also true for g_n and $g_{b,n}$. This entails $\sup_{b \in \mathbb{R}_*^d} |\frac{1}{n} \sum_{i=1}^n \{\frac{g_{b,n}(b^\top X_i)}{f_{b,n}(b^\top X_i)} - 1\} \frac{f_{b,n}(b^\top X_i)}{f_n(X_i)} - \int \{\frac{g_b(b^\top x)}{f_b(b^\top x)} - 1\} f_b(b^\top x) dx| \rightarrow 0$ a.s.

$$\begin{aligned} & \text{Indeed, we remark that } |\frac{1}{n} \sum_{i=1}^n \{\frac{g_{b,n}(b^\top X_i)}{f_{b,n}(b^\top X_i)} - 1\} \frac{f_{b,n}(b^\top X_i)}{f_n(X_i)} - \int \{\frac{g_b(b^\top x)}{f_b(b^\top x)} - 1\} f_b(b^\top x) dx| \\ &= |\frac{1}{n} \sum_{i=1}^n \{\frac{g_{b,n}(b^\top X_i)}{f_{b,n}(b^\top X_i)} - 1\} \frac{f_{b,n}(b^\top X_i)}{f_n(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{g_b(b^\top X_i)}{f_b(b^\top X_i)} - 1\} \frac{f_b(b^\top X_i)}{f(X_i)} \\ & \quad + \frac{1}{n} \sum_{i=1}^n \frac{g_b(b^\top X_i)}{f_b(b^\top X_i)} - 1\} \frac{f_b(b^\top X_i)}{f(X_i)} - \int \{\frac{g_b(b^\top x)}{f_b(b^\top x)} - 1\} f_b(b^\top x) dx| \\ &\leq |\frac{1}{n} \sum_{i=1}^n \{\frac{g_{b,n}(b^\top X_i)}{f_{b,n}(b^\top X_i)} - 1\} \frac{f_{b,n}(b^\top X_i)}{f_n(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{g_b(b^\top X_i)}{f_b(b^\top X_i)} - 1\} \frac{f_b(b^\top X_i)}{f(X_i)}| \\ & \quad + |\frac{1}{n} \sum_{i=1}^n \frac{g_b(b^\top X_i)}{f_b(b^\top X_i)} - 1\} \frac{f_b(b^\top X_i)}{f(X_i)} - \int \{\frac{g_b(b^\top x)}{f_b(b^\top x)} - 1\} f_b(b^\top x) dx| \end{aligned}$$

Moreover, since $\int |\{\frac{g_b(b^\top x)}{f_b(b^\top x)} - 1\} f_b(b^\top x)| dx \leq 2$, the law of large numbers enables us to derive:

$$|\frac{1}{n} \sum_{i=1}^n \frac{g_b(b^\top X_i)}{f_b(b^\top X_i)} - 1\} \frac{f_b(b^\top X_i)}{f(X_i)} - \int \{\frac{g_b(b^\top x)}{f_b(b^\top x)} - 1\} f_b(b^\top x) dx| \rightarrow 0 \text{ a.s..}$$

$$\begin{aligned} & \text{Moreover, } |\frac{1}{n} \sum_{i=1}^n \{\frac{g_{b,n}(b^\top X_i)}{f_{b,n}(b^\top X_i)} - 1\} \frac{f_{b,n}(b^\top X_i)}{f_n(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{g_b(b^\top X_i)}{f_b(b^\top X_i)} - 1\} \frac{f_b(b^\top X_i)}{f(X_i)}| \\ & \leq \frac{1}{n} \sum_{i=1}^n |\{\frac{g_{b,n}(b^\top X_i)}{f_{b,n}(b^\top X_i)} - 1\} \frac{f_{b,n}(b^\top X_i)}{f_n(X_i)} - \{\frac{g_b(b^\top X_i)}{f_b(b^\top X_i)} - 1\} \frac{f_b(b^\top X_i)}{f(X_i)}| \end{aligned}$$

$$\begin{aligned} & \text{and } |\{\frac{g_{b,n}(b^\top X_i)}{f_{b,n}(b^\top X_i)} - 1\} \frac{f_{b,n}(b^\top X_i)}{f_n(X_i)} - \{\frac{g_b(b^\top X_i)}{f_b(b^\top X_i)} - 1\} \frac{f_b(b^\top X_i)}{f(X_i)}| = |\frac{g_{b,n}(b^\top X_i) - f_{b,n}(b^\top X_i)}{f_n(X_i)} - \frac{g_b(b^\top X_i) - f_b(b^\top X_i)}{f(X_i)}| \\ & \leq \frac{1}{|f(X_i)| \cdot |f_n(X_i)|} \{|f(X_i)| \cdot |g_{b,n}(b^\top X_i) - g_b(b^\top X_i)| + |f(X_i) - f_n(X_i)| \cdot |g_b(b^\top X_i)| \\ & \quad + |f(X_i)| \cdot |f_{b,n}(b^\top X_i) - f_b(b^\top X_i)| + |f(X_i) - f_n(X_i)| \cdot |f_b(b^\top X_i)|\}, \end{aligned}$$

through the introduction of terms $g_b f - g_b f$ and $f f_b - f f_b$,

$$\leq \frac{O_{\mathbf{P}}(1) \cdot y_n}{\theta_n \cdot (\theta_n - y_n)} = O_{\mathbf{P}}(1) \frac{1}{\frac{y_n}{\theta_n} - \theta_n}, \text{ as a result of the very definitions of } \theta_n \text{ and } y_n \text{ respectively,}$$

$\rightarrow 0$, a.s. because, $\frac{y_n}{\theta_n^2} \rightarrow 0$ a.s., by hypothesis on θ_n .

Consequently, $\frac{1}{n} \sum_{i=1}^n |\{\frac{g_{b,n}(b^\top X_i)}{f_{b,n}(b^\top X_i)} - 1\} \frac{f_{b,n}(b^\top X_i)}{f_n(X_i)} - \{\frac{g_b(b^\top X_i)}{f_b(b^\top X_i)} - 1\} \frac{f_b(b^\top X_i)}{f(X_i)}| \rightarrow 0$, as it is a Cesàro mean.

This enables us to conclude. Similarly, we prove limits 6 and 7 page 17. \square

Proof of lemma 14.

Lemma 14 For any $p \leq d$, we have $f_{a_p}^{(p-1)} = f_{a_p}$ - see Huber's analytic method -, $g_{a_p}^{(p-1)} = g_{a_p}$ - see Huber's synthetic method - and $g_{a_p}^{(p-1)} = g_{a_p}$ - see our algorithm.

Proof :

As it is equivalent to prove either our algorithm or Huber's, we will only develop here the proof for our algorithm. Assuming, without any loss of generality, that the a_i , $i = 1, \dots, p$, are the vectors of the canonical basis, since $g^{(p-1)}(x) = g(x) \frac{f_1(x_1)}{g_1(x_1)} \frac{f_2(x_2)}{g_2(x_2)} \dots \frac{f_{p-1}(x_{p-1})}{g_{p-1}(x_{p-1})}$ we derive immediately that $g_p^{(p-1)} = g_p$.

We remark that it is sufficient to operate a change in basis on the a_i to obtain the general case. \square

Proof of lemma 15.

Lemma 15 If there exists p , $p \leq d$, such that $K(g^{(p)}, f) = 0$, then the family of $(a_i)_{i=1, \dots, p}$ - derived from the construction of $g^{(p)}$ - is free and orthogonal.

Proof :

Without any loss of generality, let us assume that $p = 2$ and that the a_i are the vectors of the canonical basis. Using a reductio ad absurdum with the hypotheses $a_1 = (1, 0, \dots, 0)$ and that $a_2 = (\alpha, 0, \dots, 0)$, where $\alpha \in \mathbb{R}$, we get $g^{(1)}(x) = g(x_2, \dots, x_d/x_1)f_1(x_1)$ and $f = g^{(2)}(x) = g(x_2, \dots, x_d/x_1)f_1(x_1)\frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)}$. Hence $f(x_2, \dots, x_d/x_1) = g(x_2, \dots, x_d/x_1)\frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)}$.

It consequently implies that $f_{\alpha a_1}(\alpha x_1) = [g^{(1)}]_{\alpha a_1}(\alpha x_1)$ since

$$1 = \int f(x_2, \dots, x_d/x_1) dx_2 \dots dx_d = \int g(x_2, \dots, x_d/x_1) dx_2 \dots dx_d \frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)} = \frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)}.$$

Therefore, $g^{(2)} = g^{(1)}$, i.e. $p = 1$ which leads to a contradiction. Hence, the family is free.

Moreover, using a reductio ad absurdum we get the orthogonality. Indeed, we have

$$\int f(x) dx = 1 \neq +\infty = \int n(a_{j+1}^\top x, \dots, a_d^\top x) h(a_1^\top x, \dots, a_j^\top x) dx. \quad \square$$

Proof of lemma 16.

Lemma 16 We have $\Theta = \{b \in \Theta \mid \int (\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)} - 1)f(x) dx < \infty\}$.

We get the result since $\int (\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)} - 1)f(x) dx = \int (\frac{g(x)f_b(b^\top x)}{g_b(b^\top x)} - f(x)) dx = 0.$ \square

Proof of propositions 11. In the same manner as in Proposition 3.4 of [BROKEZ], we prove this proposition through lemma 16. \square

Proof of propositions 4 and 9. Proposition 4 comes immediately from proposition 10 page 17 and lemma 11 page 17. Similarly, we prove proposition 9 since both $\sup_{a \in \Theta} \|\hat{b}_n(a) - a_k\|$ and $\hat{\beta}_n$ converge toward a_k a.s. in the case where f is known - see also in Appendix C, where we carry out our algorithm in the case where f is known. \square

Proof of theorem 12. Using lemma 6 page 16 and since, for any k , $g^{(k)} = g^{(k-1)} \frac{f_{a_k}}{g_{a_k}^{(k-1)}}$, we prove this theorem by induction. \square

Proof of theorems 1 and 6. We prove the theorem 1 by induction. First, by the very definition of the kernel estimator $\check{g}_n^{(0)} = g_n$ converges towards g . Moreover, the continuity of $a \mapsto f_{a,n}$ and $a \mapsto g_{a,n}$ and proposition 4 imply that $\check{g}_n^{(1)} = \check{g}_n^{(0)} \frac{f_{a,n}}{\check{g}_{a,n}^{(0)}}$ converges towards $g^{(1)}$. Finally, since, for any k , $\check{g}_n^{(k)} = \check{g}_n^{(k-1)} \frac{f_{\check{a}_k,n}}{\check{g}_{\check{a}_k,n}^{(k-1)}}$, we conclude similarly as for $\check{g}_n^{(1)}$. In a similar manner, we prove theorem 6. \square

Proof of theorem 13.

relationship (8). We consider $\Psi_j = \{\frac{f_{\check{a}_j}(\check{a}_j^\top x)}{[\check{g}^{(j-1)}]_{\check{a}_j}(\check{a}_j^\top x)} - \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}\}$. Since f and g are bounded, it is easy to prove that from a certain rank, we get, for any given x in \mathbb{R}^d

$$|\Psi_j| \leq \max(\frac{1}{[\check{g}^{(j-1)}]_{\check{a}_j}(\check{a}_j^\top x)}, \frac{1}{[g^{(j-1)}]_{a_j}(a_j^\top x)}) |f_{\check{a}_j}(\check{a}_j^\top x) - f_{a_j}(a_j^\top x)|.$$

Remark 9 First, based on what we stated earlier, for any given x and from a certain rank, there is a constant $R > 0$ independent from n , such that $\max(\frac{1}{[\check{g}^{(j-1)}]_{\check{a}_j}(\check{a}_j^\top x)}, \frac{1}{[g^{(j-1)}]_{a_j}(a_j^\top x)}) \leq R = R(x) = O(1)$. Second, since \check{a}_k is an M -estimator of a_k , its convergence rate is $O_{\mathbf{P}}(n^{-1/2})$.

Thus using simple functions, we infer an upper and lower bound for $f_{\check{a}_j}$ and for f_{a_j} . We therefore reach the following conclusion:

$$|\Psi_j| \leq O_{\mathbf{P}}(n^{-1/2}). \quad (13)$$

We finally obtain:

$$|\prod_{j=1}^k \frac{f_{\check{a}_j}(\check{a}_j^\top x)}{[\check{g}^{(j-1)}]_{\check{a}_j}(\check{a}_j^\top x)} - \prod_{j=1}^k \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}| = \prod_{j=1}^k \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)} |\prod_{j=1}^k \frac{f_{\check{a}_j}(\check{a}_j^\top x)}{[\check{g}^{(j-1)}]_{\check{a}_j}(\check{a}_j^\top x)} \frac{[g^{(j-1)}]_{a_j}(a_j^\top x)}{f_{a_j}(a_j^\top x)} - 1|.$$

Based on the relationship (13), the expression $\frac{f_{\check{a}_j}(\check{a}_j^\top x)}{[\check{g}^{(j-1)}]_{\check{a}_j}(\check{a}_j^\top x)} \frac{[g^{(j-1)}]_{a_j}(a_j^\top x)}{f_{a_j}(a_j^\top x)}$ tends towards 1 at a rate

of $O_{\mathbf{P}}(n^{-1/2})$ for all j . Consequently, $\prod_{j=1}^k \frac{f_{\tilde{a}_j}(\tilde{a}_j^\top x)}{[\tilde{g}^{(j-1)}]_{\tilde{a}_j}(\tilde{a}_j^\top x)} \frac{[g^{(j-1)}]_{a_j}(a_j^\top x)}{f_{a_j}(a_j^\top x)}$ tends towards 1 at a rate of $O_{\mathbf{P}}(n^{-1/2})$. Thus from a certain rank, we get $|\prod_{j=1}^k \frac{f_{\tilde{a}_j}(\tilde{a}_j^\top x)}{[\tilde{g}^{(j-1)}]_{\tilde{a}_j}(\tilde{a}_j^\top x)} - \prod_{j=1}^k \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}| = O_{\mathbf{P}}(n^{-1/2})O_{\mathbf{P}}(1)$. In conclusion, we obtain $|\check{g}^{(k)}(x) - g^{(k)}(x)| = g(x)|\prod_{j=1}^k \frac{f_{\tilde{a}_j}(\tilde{a}_j^\top x)}{[\tilde{g}^{(j-1)}]_{\tilde{a}_j}(\tilde{a}_j^\top x)} - \prod_{j=1}^k \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}| \leq O_{\mathbf{P}}(n^{-1/2})$.

relationship (9). The relationship 8 of theorem 13 implies that $|\frac{\check{g}^{(k)}(x)}{g^{(k)}(x)} - 1| = O_{\mathbf{P}}(n^{-1/2})$ because, for any given x , $g^{(k)}(x)|\frac{\check{g}^{(k)}(x)}{g^{(k)}(x)} - 1| = |\check{g}^{(k)}(x) - g^{(k)}(x)|$. Consequently, there exists a smooth function C of \mathbb{R}^d in \mathbb{R}^+ such that $\lim_{n \rightarrow \infty} n^{-1/2}C(x) = 0$ and $|\frac{\check{g}^{(k)}(x)}{g^{(k)}(x)} - 1| \leq n^{-1/2}C(x)$, for any x .

We then have $\int |\check{g}^{(k)}(x) - g^{(k)}(x)|dx = \int g^{(k)}(x)|\frac{\check{g}^{(k)}(x)}{g^{(k)}(x)} - 1|dx \leq \int g^{(k)}(x)C(x)n^{-1/2}dx$.

Moreover, $\sup_{x \in \mathbb{R}^d} |\check{g}^{(k)}(x) - g^{(k)}(x)| = \sup_{x \in \mathbb{R}^d} g^{(k)}(x)|\frac{\check{g}^{(k)}(x)}{g^{(k)}(x)} - 1| = \sup_{x \in \mathbb{R}^d} g^{(k)}(x)C(x)n^{-1/2} \rightarrow 0$ a.s., by theorem 12. This implies that $\sup_{x \in \mathbb{R}^d} g^{(k)}(x)C(x) < \infty$ a.s., i.e. $\sup_{x \in \mathbb{R}^d} C(x) < \infty$ a.s. since $g^{(k)}$ has been assumed to be positive and bounded - see remark 8.

Thus, $\int g^{(k)}(x)C(x)dx \leq \sup C \cdot \int g^{(k)}(x)dx = \sup C < \infty$ since $g^{(k)}$ is a density, we can therefore conclude $\int |\check{g}^{(k)}(x) - g^{(k)}(x)|dx \leq \sup C \cdot n^{-1/2} = O_{\mathbf{P}}(n^{-1/2})$. \square

relationship (10). We have

$$K(\check{g}^{(k)}, f) - K(g^{(k)}, f) = \int f(\varphi(\frac{\check{g}^{(k)}}{f}) - \varphi(\frac{g^{(k)}}{f}))dx \leq \int f S |\frac{\check{g}^{(k)}}{f} - \frac{g^{(k)}}{f}|dx = S \int |\check{g}^{(k)} - g^{(k)}|dx$$

with the line before last being derived from theorem 11 page 15 and where $\varphi : x \mapsto x \ln(x) - x + 1$ is a convex function and where $S > 0$. We get the same expression as the one found in our Proof of Relationship (9) section, we then obtain $K(\check{g}^{(k)}, f) - K(g^{(k)}, f) \leq O_{\mathbf{P}}(n^{-1/2})$. Similarly, we get $K(g^{(k)}, f) - K(\check{g}^{(k)}, f) \leq O_{\mathbf{P}}(n^{-1/2})$. We can therefore conclude. \square

Proof of lemma 17.

Lemma 17 *We keep the notations introduced in Appendix B. It holds $n = O(m^{\frac{1}{2}})$.*

Proof :

Let us first study the Huber's case. Let N be the random variable such that

$N = \sum_{j=1}^m \mathbf{1}_{\{f_m(X_j) \geq \theta_m, g(Y_j) \geq \theta_m\}}$. Since the events $\{f_m(X_j) \geq \theta_m\}$ and $\{g(Y_j) \geq \theta_m\}$ are independent from one another and since $\{g(Y_j) \geq \theta_m\} \subset \{g_m(Y_j) \geq -y_m + \theta_m\}$, we can say that

$$n = m \cdot \mathbf{P}(f_m(X_j) \geq \theta_m, g(Y_j) \geq \theta_m) \leq m \cdot \mathbf{P}(f_m(X_j) \geq \theta_m) \cdot \mathbf{P}(g_m(Y_j) \geq -y_m + \theta_m).$$

Consequently, let us study $\mathbf{P}(f_m(X_i) \geq \theta_m)$. Let $(\xi_i)_{i=1 \dots m}$ be the sequence such that, for any i and any x in \mathbb{R}^d , $\xi_i(x) = \prod_{l=1}^d \frac{1}{(2\pi)^{1/2}h_l} e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} - \int \prod_{l=1}^d \frac{1}{(2\pi)^{1/2}h_l} e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x)dx$. Hence, for any given j and conditionally to $X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_m$, the variables $(\xi_i(X_j))_{i=1 \dots m}^{i \neq j}$ are i.i.d. and centered, have the same second moment, and are such that

$$|\xi_i(X_j)| \leq \prod_{l=1}^d \frac{1}{(2\pi)^{1/2}h_l} + \prod_{l=1}^d \frac{1}{(2\pi)^{1/2}h_l} \int |f(x)|dx = 2 \cdot (2\pi)^{-d/2} \prod_{l=1}^d h_l^{-1} \text{ since } \sup_x e^{-\frac{1}{2}x^2} \leq 1.$$

Moreover, noting that $f_m(x) = \frac{1}{m} \sum_{i=1}^m \xi_i(x) + (2\pi)^{-d/2} \frac{1}{m} \sum_{i=1}^m \prod_{l=1}^d h_l^{-1} \int e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x)dx$,

we have $f_m(X_j) \geq \theta_m \Leftrightarrow \frac{1}{m} \sum_{i=1}^m \xi_i(X_j) + (2\pi)^{-d/2} \frac{1}{m} \sum_{i=1}^m \prod_{l=1}^d h_l^{-1} \int e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x)dx \geq \theta_m$

$$\Leftrightarrow \frac{1}{m-1} \sum_{\substack{i=1 \\ i \neq j}}^m \xi_i(X_j) \geq (\theta_m - (2\pi)^{-d/2} \frac{1}{m} \sum_{i=1}^m \prod_{l=1}^d h_l^{-1} \int e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x)dx - \frac{1}{m} \xi_j(X_j)) \frac{m}{m-1}$$

with $\xi_j(X_j) = 0$. Then, defining t (resp. ε) as $t = 2 \cdot (2\pi)^{-d/2} \prod_{l=1}^d h_l^{-1}$ (resp.

$\varepsilon = (\theta_m - (2\pi)^{-d/2} \prod_{l=1}^d h_l^{-1} \frac{1}{m} \sum_{i=1}^m \prod_{l=1}^d h_l^{-1} \int e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x)dx) \frac{m}{m-1}$), the Bennet's inequality -[DEVG Y85] page 160 - implies that $\mathbf{P}(\frac{1}{m-1} \sum_{\substack{i=1 \\ i \neq j}}^m \xi_i(X_j) \geq \varepsilon / X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_m) \leq 2 \cdot \exp(-\frac{(m-1)\varepsilon^2}{4t^2})$.

Finally, since the X_i are i.i.d. and since $\int (\prod_{l=1}^d e^{-\frac{1}{2}(\frac{x_l - y_l}{h_l})^2} f(x)dx) f(y)dy < 1$, then the law of large numbers implies that $\frac{1}{m} \sum_{i=1}^m \int \prod_{l=1}^d e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x)dx \rightarrow_m \int \prod_{l=1}^d e^{-\frac{1}{2}(\frac{x_l - y_l}{h_l})^2} f(x)f(y)dxdy$ a.s.

Consequently, since $0 < \nu < \frac{1}{4+d}$ - see remark 7 - and since $e^{-x} \leq x^{-\frac{1}{2}}$ when $x > 0$, we obtain, after calculation, that, from a certain rank, $\exp(-\frac{(m-1)\varepsilon^2}{4t^2}) = O(m^{-\frac{1}{4}})$, i.e., from a certain rank, $\mathbf{P}(f_m(Y_j) \geq \theta_m) = O(m^{-\frac{1}{4}})$. Similarly, we infer $\mathbf{P}(g(Y_j) \geq \theta_m) = O(m^{-\frac{1}{4}})$. In conclusion, we can say that $n = m \cdot \mathbf{P}(f_m(X_j) \geq \theta_m) \cdot \mathbf{P}(g_m(Y_j) \geq \theta_m) = O(m^{\frac{1}{2}})$. Similarly, we derive the same result as above for any step of our method as well as Huber's. \square

Proof of theorems 2 and 7. First, from lemma 13, we derive that, for any x ,

$$\sup_{a \in \mathbb{R}_*^d} |f_{a,n}(a^\top x) - f_a(a^\top x)| = O_{\mathbf{P}}(n^{-\frac{2}{4+d}}). \text{ Then, let us consider } \Psi_j = \frac{f_{a_j,n}(\tilde{a}_j^\top x)}{\tilde{g}_{a_j,n}^{(j-1)}(\tilde{a}_j^\top x)} - \frac{f_{a_j}(a_j^\top x)}{g_{a_j}^{(j-1)}(a_j^\top x)},$$

we have $\Psi_j = \frac{1}{\tilde{g}_{a_j,n}^{(j-1)}(\tilde{a}_j^\top x)g_{a_j}^{(j-1)}(a_j^\top x)} ((f_{a_j,n}(\tilde{a}_j^\top x) - f_{a_j}(a_j^\top x))g_{a_j}^{(j-1)}(a_j^\top x) + f_{a_j}(a_j^\top x)(g_{a_j}^{(j-1)}(a_j^\top x) - \tilde{g}_{a_j,n}^{(j-1)}(\tilde{a}_j^\top x)))$, i.e. $|\Psi_j| = O_{\mathbf{P}}(n^{-\frac{2}{4+d}})$ since $f_{a_j}(a_j^\top x) = O(1)$ and $g_{a_j}^{(j-1)}(a_j^\top x) = O(1)$. We can therefore conclude similarly as in theorem 13 and through lemma 17. Similarly, we derive theorem 7. \square

Proof of theorem 14. First of all, we remark that hypotheses (H'1) to (H'3) imply that $\tilde{\gamma}_n$ and $\tilde{c}_n(a_k)$ converge towards a_k in probability. Hypothesis (H'4) enables us to derive under the integrable sign after calculation, $\mathbf{P} \frac{\partial}{\partial b} M(a_k, a_k) = \mathbf{P} \frac{\partial}{\partial a} M(a_k, a_k) = 0$,

$$\mathbf{P} \frac{\partial^2}{\partial a_i \partial b_j} M(a_k, a_k) = \mathbf{P} \frac{\partial^2}{\partial b_j \partial a_i} M(a_k, a_k) = \int \varphi'' \left(\frac{g f_{a_k}}{f g_{a_k}} \right) \frac{\partial}{\partial a_i} \frac{g f_{a_k}}{f g_{a_k}} \frac{\partial}{\partial b_j} \frac{g f_{a_k}}{f g_{a_k}} f dx,$$

$$\mathbf{P} \frac{\partial^2}{\partial b_i \partial b_j} M(a_k, a_k) = - \int \varphi'' \left(\frac{g f_{a_k}}{f g_{a_k}} \right) \frac{\partial}{\partial b_i} \frac{g f_{a_k}}{f g_{a_k}} \frac{\partial}{\partial b_j} \frac{g f_{a_k}}{f g_{a_k}} f dx, \quad \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) = \int \varphi' \left(\frac{g f_{a_k}}{f g_{a_k}} \right) \frac{\partial^2}{\partial a_i \partial a_j} \frac{g f_{a_k}}{f g_{a_k}} f dx,$$

and consequently $\mathbf{P} \frac{\partial^2}{\partial b_i \partial b_j} M(a_k, a_k) = -\mathbf{P} \frac{\partial^2}{\partial a_i \partial b_j} M(a_k, a_k) = -\mathbf{P} \frac{\partial^2}{\partial b_j \partial a_i} M(a_k, a_k)$, which implies,

$$\begin{aligned} \frac{\partial^2}{\partial a_i \partial a_j} K(g \frac{f_{a_k}}{g_{a_k}}, f) &= \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) - \mathbf{P} \frac{\partial^2}{\partial b_i \partial b_j} M(a_k, a_k), \\ &= \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) + \mathbf{P} \frac{\partial^2}{\partial a_i \partial b_j} M(a_k, a_k) = \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) + \mathbf{P} \frac{\partial^2}{\partial b_j \partial a_i} M(a_k, a_k). \end{aligned}$$

The very definition of the estimators $\tilde{\gamma}_n$ and $\tilde{c}_n(a_k)$, implies that $\begin{cases} \mathbb{P}_n \frac{\partial}{\partial b} M(b, a) = 0 \\ \mathbb{P}_n \frac{\partial}{\partial a} M(b(a), a) = 0 \end{cases}$

$$\text{i.e. } \begin{cases} \mathbb{P}_n \frac{\partial}{\partial b} M(\tilde{c}_n(a_k), \tilde{\gamma}_n) = 0 \\ \mathbb{P}_n \frac{\partial}{\partial a} M(\tilde{c}_n(a_k), \tilde{\gamma}_n) + \mathbb{P}_n \frac{\partial}{\partial b} M(\tilde{c}_n(a_k), \tilde{\gamma}_n) \frac{\partial}{\partial a} \tilde{c}_n(a_k) = 0, \end{cases} \quad \text{i.e. } \begin{cases} \mathbb{P}_n \frac{\partial}{\partial b} M(\tilde{c}_n(a_k), \tilde{\gamma}_n) = 0 \text{ (E0)} \\ \mathbb{P}_n \frac{\partial}{\partial a} M(\tilde{c}_n(a_k), \tilde{\gamma}_n) = 0 \text{ (E1)} \end{cases}.$$

Under (H'5) and (H'6), and using a Taylor development of the (E0) (resp. (E1)) equation, we infer there exists $(\bar{c}_n, \bar{\gamma}_n)$ (resp. $(\tilde{c}_n, \tilde{\gamma}_n)$) on the interval $[(\tilde{c}_n(a_k), \tilde{\gamma}_n), (a_k, a_k)]$ such that

$$-\mathbb{P}_n \frac{\partial}{\partial b} M(a_k, a_k) = [(\mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k))^\top + o_{\mathbf{P}}(1), (\mathbf{P} \frac{\partial^2}{\partial a \partial b} M(a_k, a_k))^\top + o_{\mathbf{P}}(1)] a_n.$$

$$\text{(resp. } -\mathbb{P}_n \frac{\partial}{\partial a} M(a_k, a_k) = [(\mathbf{P} \frac{\partial^2}{\partial b \partial a} M(a_k, a_k))^\top + o_{\mathbf{P}}(1), (\mathbf{P} \frac{\partial^2}{\partial a^2} M(a_k, a_k))^\top + o_{\mathbf{P}}(1)] a_n)$$

with $a_n = ((\tilde{c}_n(a_k) - a_k)^\top, (\tilde{\gamma}_n - a_k)^\top)$. Thus we get

$$\begin{aligned} \sqrt{n} a_n &= \sqrt{n} \begin{bmatrix} \mathbf{P} \frac{\partial^2}{\partial b^2} M(a_k, a_k) & \mathbf{P} \frac{\partial^2}{\partial a \partial b} M(a_k, a_k) \\ \mathbf{P} \frac{\partial^2}{\partial b \partial a} M(a_k, a_k) & \mathbf{P} \frac{\partial^2}{\partial a^2} M(a_k, a_k) \end{bmatrix}^{-1} \begin{bmatrix} -\mathbb{P}_n \frac{\partial}{\partial b} M(a_k, a_k) \\ -\mathbb{P}_n \frac{\partial}{\partial a} M(a_k, a_k) \end{bmatrix} + o_{\mathbf{P}}(1) \\ &= \sqrt{n} (\mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k) \frac{\partial^2}{\partial a \partial a} K(g \frac{f_{a_k}}{g_{a_k}}, f))^{-1} \\ &\quad \cdot \begin{bmatrix} \mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k) + \frac{\partial^2}{\partial a \partial a} K(g \frac{f_{a_k}}{g_{a_k}}, f) & \mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k) \\ \mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k) & \mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k) \end{bmatrix} \cdot \begin{bmatrix} -\mathbb{P}_n \frac{\partial}{\partial b} M(a_k, a_k) \\ -\mathbb{P}_n \frac{\partial}{\partial a} M(a_k, a_k) \end{bmatrix} + o_{\mathbf{P}}(1) \end{aligned}$$

Moreover, the central limit theorem implies: $\mathbb{P}_n \frac{\partial}{\partial b} M(a_k, a_k) \xrightarrow{\mathcal{L}^{aw}} \mathcal{N}_d(0, \mathbf{P} \|\frac{\partial}{\partial b} M(a_k, a_k)\|^2)$,

$\mathbb{P}_n \frac{\partial}{\partial a} M(a_k, a_k) \xrightarrow{\mathcal{L}^{aw}} \mathcal{N}_d(0, \mathbf{P} \|\frac{\partial}{\partial a} M(a_k, a_k)\|^2)$, since $\mathbf{P} \frac{\partial}{\partial b} M(a_k, a_k) = \mathbf{P} \frac{\partial}{\partial a} M(a_k, a_k) = 0$, which leads us to the result. Finally, if f is known, we similarly prove theorem 8. \square

Proof of theorems 3 and 8. We get the theorem through proposition 10 and theorem 14. \square

Proof of proposition 12. We consider $\psi, \psi_a, \psi^{(k)}, \psi_a^{(k)}$ the characteristic functions of densities $f, f_a, g^{(k-1)}$ and $[g^{(k-1)}]_a$. We have $|\psi(ta) - \psi^{(k-1)}(ta)| = |\psi_a(t) - \psi_a^{(k-1)}(t)| \leq \int |f_a(a^\top x) - [g^{(k-1)}]_a(a^\top x)| dx$, and then $\sup_a |\psi_a(t) - \psi_a^{(k-1)}(t)| \leq \sup_a \int |f_a(a^\top x) - [g^{(k-1)}]_a(a^\top x)| dx$

$\leq \sup_a K([g^{(k-1)}]_a, f_a)$ since $\psi(ta) = \mathbb{E}(e^{ita^\top x}) = \psi_a(t)$ - where $t \in \mathbb{R}$ and $a \in \mathbb{R}_*^d$ - and since the Kullback-Leibler divergence is greater than the L^1 distance. Therefore, since, as explained in section 14 of Huber's article, we have $\lim_k K([g^{(k-1)}]_{a_k}, f_{a_k}) = 0$ we then get $\lim_k g^{(k)} = f$ - which is the Huber's representation of f . Moreover, we have $|\psi(t) - \psi^{(k)}(t)| \leq \int |f(x) - g^{(k)}(x)| dx \leq K(g^{(k)}, f)$. As explained in section 14 of Huber's article and through remark 5 page 10 as well as through the additive relationship of proposition 5, we infer that $\lim_k K(g^{(k-1)} \frac{f_{a_k}}{[g^{(k-1)}]_{a_k}}, f) = 0$. Consequently, we get $\lim_k g^{(k)} = f$ - which is our representation of f .

Proof of lemmas 1 and 2. We apply our algorithm between f and g . There exists a sequence of densities $(g^{(k)})_k$ such that $0 = K(g^{(\infty)}, f) \leq \dots \leq K(g^{(k)}, f) \leq \dots \leq K(g, f)$, (*)

where $g^{(\infty)} = \lim_k g^{(k)}$ which is a density by construction. Moreover, let $(g_n^{(k)})_k$ be the sequence of densities such that $g_n^{(k)}$ is the kernel estimate of $g^{(k)}$. Since we derive from remark 8 page 19 an integrable upper bound of $g_n^{(k)}$, for all k , which is greater than f - see also the definition of φ in the proof of theorem 4 -, then the dominated convergence theorem implies that, for any k , $\lim_n K(g_n^{(k)}, f_n) = K(g^{(k)}, f)$, i.e., from a certain given rank n_0 , we have

$$0 \leq \dots \leq K(g_n^{(\infty)}, f_n) \leq \dots \leq K(g_n^{(k)}, f_n) \leq \dots \leq K(g_n, f_n), (**)$$

Consequently, through lemma 18 page 25, there exists a k such that

$$0 \leq \dots \leq K(\Psi_{n,k}^{(\infty)}, f_n) \leq \dots \leq K(g_n^{(\infty)}, f_n) \leq \dots \leq K(\Psi_{n,k-1}^{(\infty)}, f_n) \leq \dots \leq K(g_n, f_n), (***)$$

where $\Psi_{n,k}^{(\infty)}$ is a density such that $\Psi_{n,k}^{(\infty)} = \lim_k g_n^{(k)}$. Finally, through the dominated convergence theorem and taking the limit as n in (***) we get $0 = K(g^{(\infty)}, f) = \lim_n K(g_n^{(\infty)}, f_n) \geq \lim_n K(\Psi_{n,k}^{(\infty)}, f_n) \geq 0$. The dominated convergence theorem enables us to conclude:

$$0 = \lim_n K(\Psi_{n,k}^{(\infty)}, f_n) = \lim_n \lim_k K(g_n^{(k)}, f_n). \text{ Similarly, we get lemma 2. } \square$$

Proof of lemma 18.

Lemma 18 *Keeping the notations of the proof of lemma 1, we have*

$$0 \leq \dots \leq K(\Psi_{n,k}^{(\infty)}, f_n) \leq \dots \leq K(g_n^{(\infty)}, f_n) \leq \dots \leq K(\Psi_{n,k-1}^{(\infty)}, f_n) \leq \dots \leq K(g_n, f_n), (***)$$

Proof :

First, as explained in section 4.2., we have $K(f^{(k)}, g) - K(f^{(k+1)}, g) = K(f_{a_{k+1}}^{(k)}, g_{a_{k+1}})$. Moreover, through remark 5 page 10, we also derive that $K(f^{(k)}, g) = K(g^{(k)}, f)$. Then, $K(f_{a_{k+1}}^{(k)}, g_{a_{k+1}})$ is the decreasing step of the relative entropies in (*) and leading to $0 = K(g^{(\infty)}, f)$. Similarly, the very construction of (**), implies that $K(f_{a_{k+1},n}^{(k)}, g_{a_{k+1},n})$ is the decreasing step of the relative entropies in (**) and leading to $K(g_n^{(\infty)}, f_n)$. Second, through the conclusion of the section 4.2. and the lemma 14.2 of Huber's article, we obtain that $K(f_{a_{k+1},n}^{(k)}, g_{a_{k+1},n})$ converges - decreasingly and in k - towards a positive function of n - that we will call ξ_n . Third, the convergence of $(g^{(k)})_k$ - see proposition 12 - implies that, for any given n , the sequence $(K(g_n^{(k)}, f_n))_k$ is not finite. Then, through relationship (**), there exists a k such that $0 < K(g_n^{(k-1)}, f_n) - K(g_n^{(\infty)}, f_n) < \xi_n$.

Consequently, since $Q \mapsto K(Q, P)$ is l.s.c. - see property 3 - relationship (**) implies (***). \square

Proof of theorems 4 and 9. We recall that $g_n^{(k)}$ is the kernel estimator of $\check{g}^{(k)}$. Since the Kullback-Leibler divergence is greater than the L^1 -distance, we then have $\lim_n \lim_k K(g_n^{(k)}, f_n) \geq \lim_n \lim_k \int |g_n^{(k)}(x) - f_n(x)| dx$. Moreover, the Fatou's lemma implies that

$$\begin{aligned} \lim_k \int |g_n^{(k)}(x) - f_n(x)| dx &\geq \int \lim_k [|g_n^{(k)}(x) - f_n(x)|] dx = \int |[\lim_k g_n^{(k)}(x)] - f_n(x)| dx \text{ and} \\ \lim_n \int |[\lim_k g_n^{(k)}(x)] - f_n(x)| dx &\geq \int \lim_n [|[\lim_k g_n^{(k)}] - f_n|] dx = \int |[\lim_n \lim_k g_n^{(k)}(x)] - \lim_n f_n(x)| dx. \end{aligned}$$

We then obtain that $0 = \lim_n \lim_k K(g_n^{(k)}, f_n) \geq \int |\lim_n \lim_k g_n^{(k)}(x) - \lim_n f_n(x)| dx \geq 0$, i.e. that $\int |\lim_n \lim_k g_n^{(k)}(x) - \lim_n f_n(x)| dx = 0$. Moreover, for any given k and any given n , the function $g_n^{(k)}$

is a convex combination of multivariate Gaussian distributions. As derived at remark 4, for all k , the determinant of the covariance of the random vector - with density $g^{(k)}$ - is greater than or equal to the product of a positive constant times the determinant of the covariance of the random vector with density f . The form of the kernel estimate therefore implies that there exists an integrable function φ such that, for any given k and any given n , we have $|g_n^{(k)}| \leq \varphi$. Finally, the dominated convergence theorem enables us to say that $\lim_n \lim_k g_n^{(k)} = \lim_n f_n = f$, since f_n converges towards f and since $\int |\lim_n \lim_k g_n^{(k)}(x) - \lim_n f_n(x)| dx = 0$. Similarly, we get theorem 9. \square

Proof of theorem 15. Through a Taylor development of $\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n)$ of rank 2, we get at point (a_k, a_k) : $\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) = \mathbb{P}_n M(a_k, a_k) + \mathbb{P}_n \frac{\partial}{\partial a} M(a_k, a_k)(\check{\gamma}_n - a_k)^\top + \mathbb{P}_n \frac{\partial}{\partial b} M(a_k, a_k)(\check{c}_n(a_k) - a_k)^\top + \frac{1}{2} \{ (\check{\gamma}_n - a_k)^\top \mathbb{P}_n \frac{\partial^2}{\partial a \partial a} M(a_k, a_k)(\check{\gamma}_n - a_k) + (\check{c}_n(a_k) - a_k)^\top \mathbb{P}_n \frac{\partial^2}{\partial b \partial a} M(a_k, a_k)(\check{\gamma}_n - a_k) + (\check{\gamma}_n - a_k)^\top \mathbb{P}_n \frac{\partial^2}{\partial a \partial b} M(a_k, a_k)(\check{c}_n(a_k) - a_k) + (\check{c}_n(a_k) - a_k)^\top \mathbb{P}_n \frac{\partial^2}{\partial b \partial b} M(a_k, a_k)(\check{c}_n(a_k) - a_k) \}$

Thus, lemma 10 implies $\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) = \mathbb{P}_n M(a_k, a_k) + O_{\mathbf{P}}(\frac{1}{n})$,

i.e. $\sqrt{n}(\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) - \mathbf{P} M(a_k, a_k)) = \sqrt{n}(\mathbb{P}_n M(a_k, a_k) - \mathbf{P} M(a_k, a_k)) + o_{\mathbf{P}}(1)$.

Hence $\sqrt{n}(\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) - \mathbf{P} M(a_k, a_k))$ abides by the same limit distribution as

$\sqrt{n}(\mathbb{P}_n M(a_k, a_k) - \mathbf{P} M(a_k, a_k))$, which is $\mathcal{N}(0, \text{Var}_{\mathbf{P}}(M(a_k, a_k)))$. \square

Proof of theorems 5 and 10. Through proposition 10 and theorem 15, we derive theorem 5.

Similarly, we get theorem 10. \square

References

- [AZE97] AZE D., *Eléments d'analyse convexe et variationnelle*, Ellipse, 1997.
- [BOLE] Bosq D., Lecoutre J.-P. *Livre - Theorie De L'Estimation Fonctionnelle*, Economica, 1999.
- [BROKEZ] Broniatowski M., Keziou A. *Parametric estimation and tests through divergences and the duality technique. J. Multivariate Anal. 100 (2009), no. 1, 16–36.*
- [CAMBANIS81] Cambanis, Stamatis; Huang, Steel; Simons, Gordon. *On the theory of elliptically contoured distributions. J. Multivariate Anal. 11 (1981), no. 3, 368–385.*
- [DEVGY85] Devroye, Luc; Gyrfi, Lszl. Distribution free exponential bound for the L_1 error of partitioning-estimates of a regression function. Probability and statistical decision theory, Vol. A (Bad Tatzmannsdorf, 1983), 67–76, Reidel, Dordrecht, 1985
- [DIAFREE84] Diaconis, Persi; Freedman, David. *Asymptotics of graphical projection pursuit. Ann. Statist. 12 (1984), no. 3, 793–815.*
- [DI80] Jean Dieudonné, *Calcul infinitésimal. 1980, Hermann.*
- [Frie84] Friedman, Jerome H.; Stuetzle, Werner; Schroeder, Anne. *Projection pursuit density estimation. J. Amer. Statist. Assoc. 79 (1984), no. 387, 599–608.*
- [HUBER] Huber Peter J., *Robust Statistics. Wiley, 1981 (republished in paperback, 2004)*
- [HUB85] Huber Peter J., *Projection pursuit, Ann. Statist., 13(2):435–525, 1985, With discussion.*

- [LANDS03] Landsman, Zinoviy M.; Valdez, Emiliano A. *Tail conditional expectations for elliptical distributions*. *N. Am. Actuar. J.* 7 (2003), no. 4, 55–71.
- [LIVAJ] Liese Friedrich and Vajda Igor, *Convex statistical distances, volume 95 of Teubner-Texte zur Mathematik [Teubner Texts in Mathematics]*. BSB B. G. Teubner Verlagsgesellschaft, 1987, with German, French and Russian summaries.
- [SCOTT92] Scott, David W., *Multivariate density estimation. Theory, practice, and visualization*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. A Wiley-Interscience Publication*. John Wiley and Sons, Inc., New York, 1992. xiv+317 pp. ISBN: 0-471-54770-0.
- [TOMA] Aida Toma *Optimal robust M-estimators using divergences*. *Statistics and Probability Letters*, Volume 79, Issue 1, 1 January 2009, Pages 1-5
- [VDW] van der Vaart A. W., *Asymptotic statistics, volume 3 of Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge, 1998.
- [YOHAI] Victor J. Yohai *Optimal robust estimates using the Kullback-Leibler divergence*. *Statistics and Probability Letters*, Volume 78, Issue 13, 15 September 2008, Pages 1811-1816.
- [ZMU04] Zhu, Mu. *On the forward and backward algorithms of projection pursuit*. *Ann. Statist.* 32 (2004), no. 1, 233–244.